Sentiment Analysis using Imperfect Views from Spoken Language and Acoustic Modalities

Imran Sheikh, Sri Harsha Dumpala, Rupayan Chakraborty, Sunil Kumar Kopparapu
TCS Research and Innovations-Mumbai, INDIA
{imran.as, d.harsha, rupayan.chakraborty, sunilkumar.kopparapu}@tcs.com

Abstract

Multimodal sentiment classification in practical applications may have to rely on erroneous and imperfect views, namely (a) language transcription from a speech recognizer and (b) under-performing acoustic views. This work focuses on improving the representations of these views by performing a deep canonical correlation analysis with the representations of the better performing manual transcription view. Enhanced representations of the imperfect views can be obtained even in absence of the perfect views and give an improved performance during test conditions. Evaluations on the CMU-MOSI and CMU-MOSEI datasets demonstrate the effectiveness of the proposed approach.

1 Introduction

Use of multimodal cues is especially useful for analyzing sentiment in audio-visual data like opinion videos on social media websites, call-center audio recordings etc. The different modalities, viz. language (spoken words), acoustic (speech) and visual (facial and gestures), can carry a different view of the same information like for example, sentiment. While the representations/features extracted from these individual different views add richness to the sentiment classification, the intra and inter view-interactions play an important role in better sentiment classification (Zadeh et al., 2017; Chen et al., 2018; Rajagopalan et al., 2016; Nojavanasghari et al., 2016; Xu et al., 2013).

Although fusion of multi-view information is being extensively explored, the challenges associated with the presence of noise and irregularities in a view has received very less attention. For instance, multimodal sentiment classification systems have typically used manual, and hence, error free language transcriptions and exploited the interaction of other views with this noise free language view (Zadeh et al., 2018, 2017). However, a practical system will have to rely on a language transcription from an Automatic Speech Recognition (ASR) engine, which is inherently prone to errors due to ambient/channel noises in acoustic environments (Gong, 1995; Li et al., 2014), language domain mismatch, emotion in speech (Athanaselis et al., 2005), etc. Similarly, existing and popularly used representations of the acoustic view have generally under-performed compared to the language view (Poria et al., 2017; Zadeh et al., 2018; Pérez-Rosas et al., 2013), indicating that the acoustic view or its representations, by themselves, may not be discriminative enough for robust sentiment classification.

Assuming the ASR (language transcription) and acoustic views as imperfect views, the focus of this work is on improving the representations of these noisy views, riding on the representations of the better performing view. We show that the representations obtained from automatic transcriptions of spoken language and those from the acoustic views can be enhanced using corresponding representations from manual transcriptions of spoken language. Enhanced representations of the imperfect views can be obtained even in absence of the perfect views during test conditions. Deep canonical correlation analysis (DCCA) (Andrew et al., 2013) is used to improve the representations of the imperfect views. The rest of the paper is organized as follows. Section 2 describes a method to improve imperfect or erroneous views. Section 3 presents the different components in our multimodal sentiment classification system. Experiments are discussed in Section 4 followed by a discussion on results and conclusion in Section 5.

Proceedings of the First Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML), pages 35–39,
Melbourne, Australia July 20, 2018. ©2018 Association for Computational Linguistics

2 Improving representations of spoken language and acoustic views

Multimodal sentiment classification works have mainly relied on the manual transcription of the spoken utterances. In a practical and real life scenario, the text transcriptions are not readily available and are required to be obtained from an ASR engine. While ASR systems have seen large improvements with the use of deep learning methods, their performance is impacted by mismatched train-test conditions. As a result, practical multimodal sentiment classification systems will have to rely on imperfect spoken language views.

On the other hand, acoustic views used by multimodal sentiment classification systems have shown poor performance compared to that of the language view. This might indicate that either the acoustic view or its utterance level audio representations are not discriminative enough for sentiment classification. Recent classification models capture interactions across view/modality and produce better sentiment classification results (Zadeh et al., 2018, 2017). In contrast to this, our work focuses on improving representations of the imperfect views using representations of the better performing view. Utterance level representations obtained from ASR view and the acoustic view are improved using the representations extracted from manual transcriptions of spoken language. These representation improvements are achieved using DCCA (Andrew et al., 2013).

2.1 Deep canonical correlation analysis

Given the representations of two different views of the same signal, DCCA learns a pair of nonlinear transformations such that the transformed representations for the two views are maximally correlated. The individual transformed representations from a DCCA model have been shown to capture information from both the views and as a result outperform the original individual representations (Andrew et al., 2013; Wang et al., 2015; Shao et al., 2015). Figure 1 shows a high level block representation of DCCA. (f_{v1}, f_{v2}) are representations of two views of the same input data, nonlinear transformations are carried out using DNNs and canonical correlations are computed on the DNN transformed representations ($\hat{f}_{v1}$, $\hat{f}_{v2}$). During training, representations for the two views are extracted from the train set and used to train the DNN's such that canonical correlation between the transformed representations is maximized. Thus, the goal is to learn parameters W_1^*, W_2^* for DNN_{View1}, DNN_{View2}, such that:

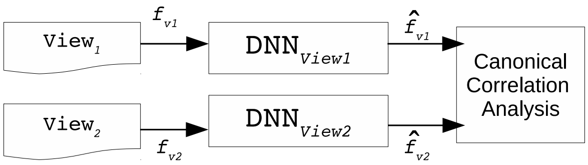

Figure 1: Improving views using DCCA.

$$(W_1^*, W_2^*) = \underset{W_1, W_2}{argmax} \;\; corr(g_1(f_{v1}; W_1), \\ g_2(f_{v2}; W_2))$$

$$\hat{f}_{v1} = g_1(f_{v1}; W_1) \;\; , \;\; \hat{f}_{v2} = g_2(f_{v2}; W_2)$$

where, g_1, g_2 denote the nonlinear transformations of DNN_{View1} and DNN_{View2} respectively. Once the DNNs are trained they are used to obtain the transformed or enhanced representations.

3 Sentiment classification using language and acoustic views

This section describes our complete system for sentiment classification which uses language and acoustic views. We first discuss the views and their representations and then describe the method adopted to fuse and classify these representations.

3.1 Spoken language views & representations

3.1.1 Manual and ASR views

A typical view of the spoken language modality is the word level manual transcription of the spoken utterances. However, in a realistic scenario manual transcriptions are not available and the system has to rely on automatic transcriptions of the spoken language. Therefore, we consider the automatic transcriptions from a general purpose ASR engine as a practical spoken language view.

To obtain the ASR view, we use the public domain Kaldi ASR toolkit (Povey et. al., 2011) along with the ASpIRE Chain acoustic models (Peddinti et al., 2015; Povey, 2017). The accompanying pretrained language model is used as it is. When evaluated on the 2199 speech utterances in the CMU-MOSI dataset (Zadeh et al., 2016), this ASR setup gives a mean word error rate of 49.2% (with a standard deviation of 32.0). Its performance in terms of correctly recognized words is 66.8%.

3.1.2 CNN based representation

Representations for the spoken language views are obtained using a text convolutional neural network (CNN) (Kim, 2014). Each utterance is represented as the concatenation of 300-dimensional GloVe embeddings (Pennington et al., 2014). Then 1-dimensional convolution kernels are applied to the concatenated word embeddings. The CNN has two convolutional layers, with the first layer having two kernels of size 3 and 4 with 50 feature maps each and the second layer having a kernel of size 2 with 100 feature maps. Each convolutional layer was followed by a 2×2 max pooling layer. A fully connected layer transforms the CNN extracted features into a 300-dimensional vector.

3.2 Representation of acoustic view

As a representation of the acoustic view, we extract a large set of high level descriptors (HLDs) from low level audio descriptors (LLDs) like voice probability, MFCCs, pitch, RMS energies and their delta regression coefficients. Since the HLDs are (up to fourth order) statistics of LLDs extracted over smaller (20 ms) frames, the dimension of the acoustic features remain same (i.e. 384) for all utterances. We used the $IS09$ configuration from the openSMILE toolkit (Eyben et al., 2009).

3.3 Fusion and sentiment classification

Bi-modal representations for utterance level sentiment classification are obtained by first extracting the representations of (manually transcribed and ASR) spoken language views and those for the acoustic view, as discussed in Section 3.1. Then representations of automatically transcribed spoken language view and those for the acoustic view are improved using DCCA, as discussed in Section 2.1. Finally the improved representations are concatenated to obtain a bi-modal representation.

We use a bi-directional LSTM-RNN to label utterance level sentiments based on the bi-modal representations. Sequence labeling with LSTM-RNNs can account for contextual information from adjacent inputs as well as the overall input sequence and has been shown to perform better on several tasks (Graves and Schmidhuber, 2005; Graves et al., 2008; Poria et al., 2017; Sheikh et al., 2017). Let us denote the bi-modal representations as $(x_1, ...x_{t-1}, x_t, x_{t+1}..., x_N)$, where x_t represents the current utterance and N is the number of utterances in a video. We followed the hierarchical training discussed in (Poria et al., 2017). Each bi-modal representation (x_t) is input to the forward and backward LSTM-RNNs to obtain the hidden layer activations h_t^F and h_t^B. These concatenated activations (c_t) are fed to softmax classifier,

$$p_t(i) = \frac{exp(c_{ti}.W_C + b_C)}{\sum_j exp(c_{tj}.W_C + b_C)} \qquad (1)$$

where $p_t(i)$ denotes the posterior probability of output class i for utterance at t; W_C and b_C are weight and bias parameters of the softmax layer.

4 Experiments and results

4.1 Datasets

We present our results and analysis on two datasets, namely, (a) CMU-MOSI (Zadeh et al., 2016) and (b) CMU-MOSEI (Zadeh, 2018a). CMU-MOSI consists of 93 movie related opinion videos from YouTube, segmented into 2199 clips/utterances. CMU-MOSEI consists of about 2500 multi-domain monologue videos from YouTube, segmented into $23,500$ clips/utterances.

Both CMU-MOSI and CMU-MOSEI datasets are annotated with utterance level sentiment labels in the range $[-3, 3]$. We focus on binary sentiment classification in which labels $[-3, 0]$ are considered as negative and $[1, 3]$ are considered as positive sentiments. For CMU-MOSI, we used the train, validation and test split provided by the CMU Multimodal Data SDK (Zadeh, 2018b). The SDK also provides a train, validation and test split for CMU-MOSEI. However, as the test set labels were not available at the time of submission of this paper, we treated 200 videos from the original validation set as our test set. The remaining 100 videos from the original validation set and an additional 150 videos from the original train set are considered as our validation set.

4.2 Experiment setup

We evaluate the performance of the spoken language and acoustic views, individually and in combination. The manual and ASR transcriptions of the language view are denoted as MT and AT, respectively. The acoustic view is denoted as AU. Enhanced (representations of) ASR view and acoustic view are denoted as AT$_\uparrow$ and AU$_\uparrow$, respectively. They were enhanced using (representations of) the manual transcription view, using the DCCA model described in Section 2.1. Our DCCA models use DNNs with 3 hidden layers and sigmoids.

4.3 Sentiment classification results

Table 1 presents the % accuracy (Acc.) and F-score (F1) for binary sentiment classification on the CMU-MOSI and CMU-MOSEI datasets. The results are divided into four sections, viz. (I) the 'ideal' baseline results achieved by the LSTM-RNN classifier on the manual transcription and acoustic views, (II) the 'practical' baseline results achieved with the imperfect ASR view, (III) the results obtained, for the practical scenario, by the proposed approach with DCCA enhanced views and (IV) the improvement on using DCCA enhanced acoustic view with manual transcriptions.

Table 1: Sentiment classification performance using a bi-directional LSTM-RNN classifier.

		MOSI		MOSEI	
		Acc.	F1	Acc.	F1
I	AU	50.6	50.0	59.4	58.0
	MT	73.5	73.1	68.7	68.6
	MT+AU	71.4	71.0	68.7	68.7
II	AT	69.1	68.6	68.0	67.5
	AT+AU	69.4	69.2	68.1	67.9
III	AU$_\uparrow$	51.6	51.1	58.9	59.3
	AT$_\uparrow$	70.2	69.7	68.8	68.7
	AT$_\uparrow$+AU$_\uparrow$	**70.9**	**70.7**	**69.1**	**69.0**
IV	MT+AU$_\uparrow$	**74.6**	**74.1**	**69.4**	**69.3**

5 Discussion

5.1 Performance of ASR view (AT)

Comparison of MT and AT views in sections I and II of Table 1 shows that the AT view degrades the classification performance Accuracy and F-score reduce by 4.4% and 4.5% absolute for CMU-MOSI and by 0.6% and 0.8% absolute for CMU-MOSEI[1]. Similarly, degradations are also present in the bimodal setup (MT+AU vs AT+AU).

5.2 Performance of acoustic view (AU)

The acoustic view (AU) in itself gives a poor performance for CMU-MOSI and a relatively better performance for CMU-MOSEI. However, when fused along with the language views (MT or AT), it results in small or no improvement and sometimes a degradation. This indicates that the raw acoustic views or its existing representations may not always contribute for sentiment classification, due to the existence of encoded and decoded sentiments as discussed in (Chakraborty et al., 2018).

5.3 Improvements with DCCA

As discussed above, the ASR and acoustic views (AT and AU) reduced the classification scores. Section III of Table 1 shows that our approach to enhance the imperfect views using DCCA can lead to significant improvements. ASR view (AT vs AT$_\uparrow$) F-scores improve by 1.1% (CMU-MOSI) and 1.2% (CMU-MOSEI) absolute. Acoustic view (AU vs AU$_\uparrow$) F-scores improve by 1.1% (CMU-MOSI) and 1.3% (CMU-MOSEI) absolute. F-scores for the bimodal system with ASR view (AT+AU vs AT$_\uparrow$+AU$_\uparrow$) improve by 1.5% (CMU-MOSI) and 1.1% (CMU-MOSEI) absolute. Bimodal system with manual transcription and DCCA enhanced acoustic view (MT+AU vs MT+AU$_\uparrow$) also shows F-score improvements, of 3.1% (CMU-MOSI) and 0.6% (CMU-MOSEI).

5.4 ASR view improvements with non contextual classifier

As discussed in (Poria et al., 2017), the bi-directional LSTM-RNN exploits contextual information from the adjacent utterances and the entire video. In order to obtain the improvements due to DCCA alone we evaluated the performances of MT, AT and AT$_\uparrow$ with a non contextual classifier. We trained logistic regression models which classify the utterance level CNN representations independently into positive and negative sentiments. Table 2 reports the resulting % accuracies. ASR view (AT vs AT$_\uparrow$) accuracies improve by 1.4% and 1.9% absolute due to DCCA.

Table 2: Improvement in ASR view accuracy using a non contextual classifier.

	MOSI	MOSEI
MT	71.1	67.5
AT	63.7	63.8
AT$_\uparrow$	65.1	65.7

6 Conclusion

Erroneous ASR views and weak acoustic views of videos can degrade sentiment classification performance in practical scenarios. We observed degradations (up to 4.5% absolute) in F-score on standard CMU-MOSI dataset, using a popular ASR setup and an utterance level contextual LSTM-RNN classifier The effect could be more severe on multimodal systems relying on word level fusion. Our approach to improve the imperfect views using canonical correlation analysis shows significant improvements (up to 3.1% absolute).

[1] We found that manual transcriptions of several utterances in the CMU-MOSEI dataset are unreliable and hence its performance of MT would be higher than that obtained.

References

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *Proceedings of the International Conference on Machine Learning*, volume 28, pages 1247–1255.

T. Athanaselis, S. Bakamidis, I. Dologlou, R. Cowie, E. Douglas-Cowie, and C. Cox. 2005. Asr for emotional speech: Clarifying the issues and enhancing performance. *Neural Netw.*, 18(4):437–444.

Rupayan Chakraborty, Meghna Pandharipande, and Sunil Kumar Kopparapu. 2018. *Analyzing Emotion in Spontaneous Speech*, 1st edition. Springer Publishing Company, Incorporated.

Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrusaitis, Amir Zadeh, and Louis-Philippe Morency. 2018. Multimodal sentiment analysis with word-level fusion and reinforcement learning. *CoRR*, abs/1802.00924.

Florian Eyben, Felix Weninger, Martin Woellmer, and Bjoern Schuller. 2009. openSMILE. `http://www.audeering.com/research/opensmile`. Accessed: 2017.

Yifan Gong. 1995. Speech recognition in noisy environments: A survey. *Speech Commun.*, 16(3):261–291.

A. Graves and J. Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm networks. In *Proceedings of International Joint Conference on Neural Networks*, pages 2047–2052.

Alex Graves, Marcus Liwicki, Horst Bunke, Jürgen Schmidhuber, and Santiago Fernández. 2008. Unconstrained on-line handwriting recognition with recurrent neural networks. In *Advances in Neural Information Processing Systems 20*, pages 577–584.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*, pages 1746–1751.

J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach. 2014. An overview of noise-robust automatic speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 22(4):745–777.

Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltrušaitis, and Louis-Philippe Morency. 2016. Deep multimodal fusion for persuasiveness prediction. In *ICMI*, pages 284–288.

V. Peddinti, G. Chen, V. Manohar, T. Ko, D. Povey, and S. Khudanpur. 2015. Jhu aspire system: Robust lvcsr with tdnns, ivector adaptation and rnn-lms. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 539–546.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Verónica Pérez-Rosas, Rada Mihalcea, and Louis Philippe Morency. 2013. Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 973–982.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 873–883.

Daniel Povey. 2017. Kaldi models. http://kaldi-asr.org/models.html.

Daniel Povey et. al. 2011. The kaldi speech recognition toolkit. In *IEEE Workshop on Automatic Speech Recognition and Understanding*.

Shyam Sundar Rajagopalan, Louis-Philippe Morency, Tadas Baltrusaitis, and Roland Goecke. 2016. Extending long short-term memory for multi-view structured learning. In *Computer Vision – ECCV 2016*, pages 338–353.

J. Shao, Z. Zhao, F. Su, and T. Yue. 2015. 3view deep canonical correlation analysis for cross-modal retrieval. In *2015 Visual Communications and Image Processing (VCIP)*, pages 1–4.

I. Sheikh, D. Fohr, and I. Illina. 2017. Topic segmentation in asr transcripts using bidirectional rnns for change detection. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 512–518.

W. Wang, R. Arora, K. Livescu, and J. A. Bilmes. 2015. Unsupervised learning of acoustic features via deep canonical correlation analysis. In *ICASSP*, pages 4590–4594.

Chang Xu, Dacheng Tao, and Chao Xu. 2013. A survey on multi-view learning. *CoRR*, abs/1304.5634.

Amir Zadeh. 2018a. CMU-MOSEI dataset. `http://multicomp.cs.cmu.edu/resources/cmu-mosei-dataset/`. Accessed: 2018.

Amir Zadeh. 2018b. CMU Multimodal Data SDK. `https://github.com/A2Zadeh/CMU-MultimodalDataSDK`. Accessed: 2018.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *CoRR*, abs/1707.07250.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. *CoRR*, abs/1802.00927.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *CoRR*, abs/1606.06259.