Robust-MSA: Understanding the Impact of Modality Noise on Multimodal Sentiment Analysis

Huisheng Mao[1,2], **Baozheng Zhang**[1,3], **Hua Xu**[1,2*], **Ziqi Yuan**[1,2], **Yihe Liu**[1,3]

[1] State Key Laboratory of Intelligent Technology and Systems, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[2] Beijing National Research Center for Information Science and Technology(BNRist), Beijing 100084, China
[3] School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang 050018, China
xuhua@mail.tsinghua.edu.cn

Abstract

Improving model robustness against potential modality noise, as an essential step for adapting multimodal models to real-world applications, has received increasing attention among researchers. For Multimodal Sentiment Analysis (MSA), there is also a debate on whether multimodal models are more effective against noisy features than unimodal ones. Stressing on intuitive illustration and in-depth analysis of these concerns, we present Robust-MSA, an interactive platform that visualizes the impact of modality noise as well as simple defence methods to help researchers know better about how their models perform with imperfect real-world data.

Introduction

Multimodal Sentiment Analysis (MSA) is an increasingly popular task in multimodal machine learning (Baltrušaitis, Ahuja, and Morency 2018; Soleymani et al. 2017). It aims to analyze speaker's sentiment from a short video clip containing three modalities: visual, audio and text. Although researchers have achieved promising improvements over the years (Rahman et al. 2020; Hazarika, Zimmermann, and Poria 2020), models of this task are not as widely used in applications as those of other popular machine learning tasks. Lacking of ability to give correct predictions on real-world samples is a major cause. Videos in popular MSA datasets, such as MOSI (Zadeh et al. 2016), MOSEI (Zadeh et al. 2018) and CH-SIMS (Yu et al. 2020; Liu et al. 2022), are usually handpicked samples: speakers' faces are frontal without occlusion; their voices are clear without noise or interruption; the text transcripts are manually revised thus have minimal error. In real-world scenarios, however, such "perfect" samples are not the common case. Speakers may turn away from the camera; their voices may be overwhelmed by environment noise; text transcript, the dominant modality, has to be obtained via Automatic Speech Recognition (ASR) and thus may have devastating errors.

文本通常经过精心挑选，与real-world差异较大

To address these problems, researchers have identified a key challenge in MSA: how to effectively improve model robustness against modality noise (Liang, Zadeh, and Morency 2022; Li et al. 2020). In order to develop a robust model, it's essential to understand how modality noise affect existing models. In this paper, we present Robust-MSA, an interactive visualization platform for understanding what kind of influence modality noise impose on MSA models.

* Hua Xu is the corresponding author.

Demonstrating Robust-MSA

Robust-MSA takes user-generated videos as input. Speech recognition is proceeded automatically after uploading the video. Manually revising of the generated transcript is needed for obtaining a "perfect" instance. Robust-MSA then aligns video with the transcript, and offers customization of noise on word granularity. The platform visualizes video-text alignment results for both original and noise-injected version of the video, and highlights how they differ and lead to wrong predictions. Furthermore, Robust-MSA provides visualization of modality features in a timeline view. This helps researchers better understand how noise affect the feature extraction process and lead to mispredictions.

Noise Generation

Modality noise in MSA usually results in several common problems at feature level. For example, occlusion and bad camera angle may cause facial detection failure, which leads to zero values in corresponding feature dimension; noisy environment and bad microphone reception may result in ineffective audio features; ASR algorithm and typos may introduce transcript errors and further lead to incorrect text features. Robust-MSA provides six different noise simulator imitating real-world data imperfections on the word granularity. For video modality, "Blank-Screen" and "Gaussian-Blur" are supported. For audio modality, the platform provides "Mute" method and six different kind of "additive background noise" from DEMAND dataset (Thiemann, Ito, and Vincent 2013). For text modality, the options are "Replace" and "Remove". These six methods can well simulate most modality noises from real-world scenarios since they result in the same problems at feature level.

As shown in Figure 1A, to add modality noise to a video, simply drag one of the six methods, drop it onto a word, and the method will be automatically applied to the corresponding modality of the word. The added noise will be highlighted with different background colors according to their modality.

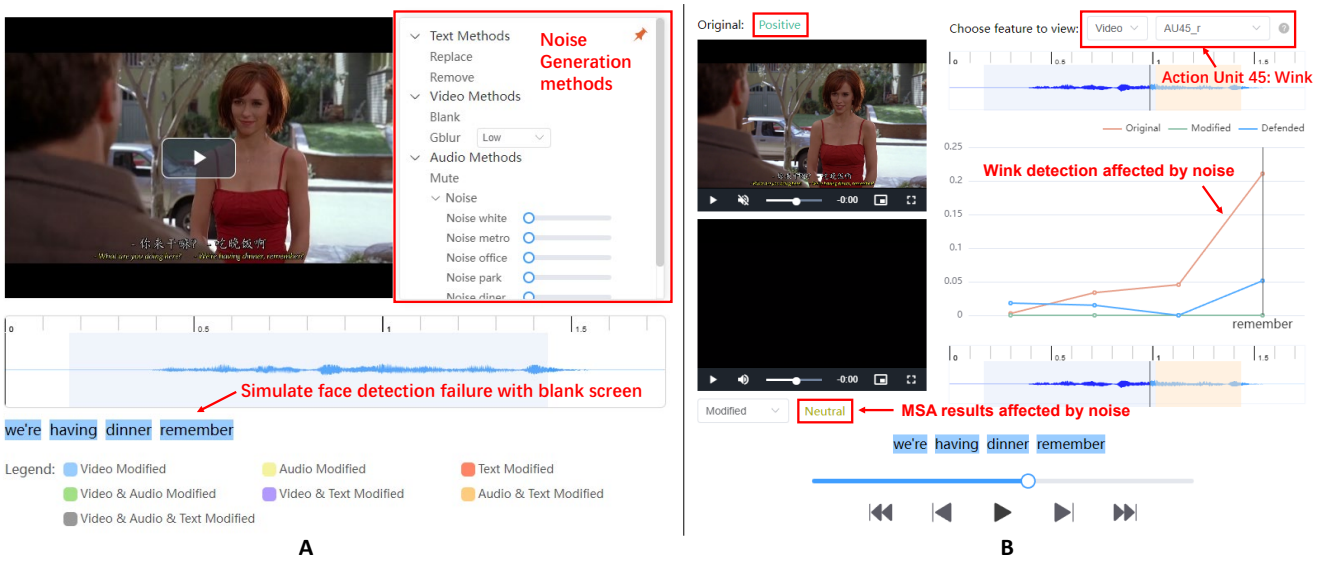

Figure 1: Noise Influence Demonstration. Left: Noise Injection, Right: feature and prediction comparison.

Noise Defence Methods

Robust-MSA provides three simple noise defence methods, including audio denoising, video motion compensated interpolation (MCI) at raw data level, and feature interpolation at feature level. The audio denoising method denoises the raw audio wave with Fast Fourier Transform (FFT); the video MCI method generates missing frames with Enhanced predictive zonal search algorithm (EPZS) algorithm (Tourapis 2002); the feature interpolation simply does a linear interpolation on missing features. The defended video and features can be viewed on final result page. Researchers can experiment for themselves to figure out whether these simple defence methods help to improve model robustness or not.

End-to-End MSA Pipeline

Feature Extraction. For real-time application, identical modality features for both training and inference stage are required. Specifically, eGeMapsv02 (Eyben et al. 2015) feature set is adopted as acoustic features, facial landmarks (Zadeh et al. 2017b) and Action Units (AU) (Baltrušaitis, Mahmoud, and Robinson 2015) are extracted as visual features, BERT (Devlin et al. 2018) language model is selected to process textual features. Moreover, a pretrained Wav2vec2 model (Baevski et al. 2020) is used to generate timestamps for video/audio to text alignment. All above customized feature extraction are performed with the help of MMSA-FET toolkit (Mao et al. 2022).

Integrated MSA Models. Currently, Robust-MSA supports eight MSA benchmark models including TFN (Zadeh et al. 2017a), LMF (Liu et al. 2018), MISA (Hazarika, Zimmermann, and Poria 2020), MAG-BERT (Rahman et al. 2020), Self-MM (Yu et al. 2021), MMIM (Han, Chen, and Poria 2021) and TFR-Net (Yuan et al. 2021) for performances comparison on the noisy environment. All models are trained on MOSEI (Zadeh et al. 2018). The final sentiment prediction shown in Figure 1B, Robust-MSA averages the models' outputs and map the score into three classes, "Negative", "Neutral" and "Positive".

Noise Influence Demonstration

To help researchers better understand the affect of modality noise on both extracted features and sentiment predictions, Robust-MSA presents the original video, its noised-injected version, and noise-defended version in alignment with the transcript. To show the alignment results, corresponding audio segment and text are highlighted accompanied by video. Users can also click on a word or use the control buttons below to quickly navigate through words in video and audio. With these convenient operations, users can easily pinpoint video and audio segments where simulated modality noise is introduced. Moreover, the platform visualizes modality features of three versions of the video in a line chart where the x-axis represents corresponding words in the transcript, as shown in the right of Figure 1.

Engaging the Audience

Our demonstration focus on how even a tiny inconspicuous modality noise, such as facial occlusion in a few frames, may lead to incorrect predictions even for models designed to overcome such noises in MSA tasks. As shown in Figure 1A, we modified the original video by dropping the entire visual modality to simulate face detection failure. The results are shown in Figure 1B, the noise-injected video is classified as "Neutral" while the original one is "Positive". From feature view we can inspect the "wink" action unit, which is a crucial visual cue for sentiment prediction.

Hopefully, the demonstration will raise more concerns on this topic and help the audience realize the importance of model robustness to applying MSA models in real-world applications.

Acknowledgments

This paper is founded by National Natural Science Foundation of China (Grant No. 62173195) and Beijing Academy of Artificial Intelligence(BAAI).

References

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33: 12449–12460.

Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443.

Baltrušaitis, T.; Mahmoud, M.; and Robinson, P. 2015. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 6, 1–6. IEEE.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Eyben, F.; Scherer, K. R.; Schuller, B. W.; Sundberg, J.; André, E.; Busso, C.; Devillers, L. Y.; Epps, J.; Laukka, P.; Narayanan, S. S.; et al. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2): 190–202.

Han, W.; Chen, H.; and Poria, S. 2021. Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 9180–9192.

Hazarika, D.; Zimmermann, R.; and Poria, S. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, 1122–1131.

Li, B.; Li, C.; Duan, F.; Zheng, N.; and Zhao, Q. 2020. Tpfn: Applying outer product along time to multimodal sentiment analysis fusion on incomplete data. In *European Conference on Computer Vision*, 431–447. Springer.

Liang, P. P.; Zadeh, A.; and Morency, L.-P. 2022. Foundations and Recent Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions. *arXiv preprint arXiv:2209.03430*.

Liu, Y.; Yuan, Z.; Mao, H.; Liang, Z.; Yang, W.; Qiu, Y.; Cheng, T.; Li, X.; Xu, H.; and Gao, K. 2022. Make Acoustic and Visual Cues Matter: CH-SIMS v2. 0 Dataset and AV-Mixup Consistent Module. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION*, 247–258.

Liu, Z.; Shen, Y.; Lakshminarasimhan, V. B.; Liang, P. P.; Zadeh, A. B.; and Morency, L.-P. 2018. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2247–2256.

Mao, H.; Yuan, Z.; Xu, H.; Yu, W.; Liu, Y.; and Gao, K. 2022. M-SENA: An Integrated Platform for Multimodal Sentiment Analysis. *arXiv preprint arXiv:2203.12441*.

Rahman, W.; Hasan, M. K.; Lee, S.; Zadeh, A. B.; Mao, C.; Morency, L.-P.; and Hoque, E. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2359–2369.

Soleymani, M.; Garcia, D.; Jou, B.; Schuller, B.; Chang, S.-F.; and Pantic, M. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65: 3–14.

Thiemann, J.; Ito, N.; and Vincent, E. 2013. The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings. In *Proceedings of Meetings on Acoustics ICA2013*, volume 19, 035081. Acoustical Society of America.

Tourapis, A. M. 2002. Enhanced predictive zonal search for single and multiple frame motion estimation. In *Visual Communications and Image Processing 2002*, volume 4671, 1069–1079. SPIE.

Yu, W.; Xu, H.; Meng, F.; Zhu, Y.; Ma, Y.; Wu, J.; Zou, J.; and Yang, K. 2020. CH-SIMS: A Chinese Multimodal Sentiment Analysis Dataset with Fine-grained Annotation of Modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3718–3727. Online: Association for Computational Linguistics.

Yu, W.; Xu, H.; Yuan, Z.; and Wu, J. 2021. Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning for Multimodal Sentiment Analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 10790–10797.

Yuan, Z.; Li, W.; Xu, H.; and Yu, W. 2021. Transformer-based Feature Reconstruction Network for Robust Multimodal Sentiment Analysis. In *Proceedings of the 29th ACM International Conference on Multimedia*, 4400–4407.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017a. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1103–1114.

Zadeh, A.; Chong Lim, Y.; Baltrusaitis, T.; and Morency, L.-P. 2017b. Convolutional experts constrained local model for 3d facial landmark detection. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2519–2528.

Zadeh, A.; Zellers, R.; Pincus, E.; and Morency, L.-P. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6): 82–88.

Zadeh, A. B.; Liang, P. P.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2236–2246.