Analyzing Modality Robustness in Multimodal Sentiment Analysis

Anonymous ACL submission

Abstract

Building robust multimodal models are crucial to achieving reliable deployment in the wild. Despite its importance, less attention has been paid to identifying and improving the robustness of Multimodal Sentiment Analysis (MSA) models. In this work, we hope to address that by (i) Proposing simple diagnostic checks for modality robustness in a trained multimodal model. Using these checks, we find MSA models to be highly sensitive to a single modality, which creates issues in their robustness; (ii) We analyze well-known robust training strategies to alleviate the issues. Critically, we observe that robustness can be achieved *without* compromising on the original performance. We hope our extensive study–performed across five models and two benchmark datasets–and proposed procedures would make robustness an integral component in MSA research. Our diagnostic checks and robust training solutions are simple to implement and shall be released at https://github.com/XXXX

1 Introduction

Multimodal Sentiment Analysis (MSA) is a burgeoning field of research that has seen accelerated developments in recent years. Numerous models have been proposed that utilize multiple modalities such as audio, visual, and language signals to predict sentiments, emotions, and other forms of affect. While progress in MSA has been driven mainly by improvements in multimodal performance, we call for attention towards an equally important aspect in multimodal systems – *multimodal robustness*. Robustness is crucial when models are deployed in the wild, where it is common to encounter inadvertent errors in the source modalities due to data loss, data corruption, jitter, privacy issues, amongst others.

A well-known fact in the MSA research is that *language* modality tends to be the most effective, which has prompted models to utilize language as its core modality (Wu et al., 2021; Han et al., 2021a; Zeng et al., 2021). In this work, we focus on skewed dependence on language and try to understand how it affects the robustness of MSA models. Specifically, we ask,

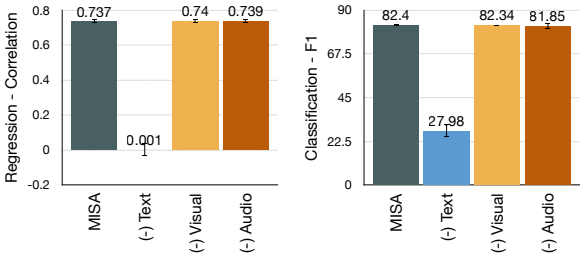

Figure 1: Removing modalities one at a time from the testing set of CMU-MOSI (Zadeh et al., 2016) on a trained MISA (Hazarika et al., 2020).

RQ1: *Are models in MSA over-reliant on a subset of modalities, particularly language?*

RQ2: *If yes, what implications does it have on modality robustness?*

To answer **RQ1**, we look at Fig. 1. The figure illustrates a setup where we fully remove one modality during testing on the MISA model (Hazarika et al., 2020). Here, we observe a sharp drop in performance when language modality is removed but do not see statistically significant drops when audio or visual modalities are removed. This observation aligns with recent findings in the MSA literature highlighting the dominance of language.

This brings us to **RQ2** where we try to understand the robustness implications over this dominance. We design an elaborate study in § 3—over five state-of-the-art (SOTA) MSA models and across two benchmark datasets—where we propose diagnostic checks to understand *modality robustness*, i.e., how robust are models against modality errors such as missing or noisy modalities.

Based on our findings, we then proceed to ask,

RQ3: *How can we improve the robustness of these models?*

RQ4: *Does robust training lead to a performance trade-off?*

For **RQ3**, we study well-known robust training methods, that act as a pre-emptive strategy to reduce the performance drops. Critically, our training is *model-agnostic* and can be easily included in any existing multimodal model (§ 4). For **RQ4**, we observe that our method to improve robustness *does not* trade-off with the final performance on the clean testing set, thus achieving similar performance as the original model.

2 Related Works

While MSA has received increased attention in recent times, the topic of robustness has not taken center stage. Fortunately, few works have started changing this trend. (Gat et al., 2020) reveals how multimodal classifiers often utilize a subset of modalities, which they addressed by inducing uniform contribution from all input modalities. In MSA, multiple works over-rely on language modality to improve the performance. (Wu et al., 2021) constructs a text-centered shared-private framework for multimodal fusion, and (Han et al., 2021a) obtains two text-related modal pairs and iteratively push the interaction between modalities to supplement information for better performance. While this has enabled performance boosts, our goal is to explore the double-edged nature of this feature and how it impacts robustness. Our motivation for diagnostics is similar to (Frank et al., 2021), but unlike them, we do not perturb the raw data (such as image patches). Instead, we intervene on modality representations, which is easier to integrate with existing models and do not require prior knowledge of the modality structure.

To address robustness in MSA, (Tsai et al., 2018) proposes a factored model that can accommodate modality drops. Also, (Ma et al., 2021) introduces modality drops during training and testing and uses meta-learning to make models robust. However, our work comprises some crucial distinctions: i) Unlike these works, our diagnostics and robust training do not require sophisticated architecture and can be easily integrated into existing models. ii) We perform an exhaustive analysis of robustness across multiple models, which is previously not done in the MSA literature.

3 Testing Robustness via Diagnostic Checks

In this section we perform an elaborate study on modality robustness by simulating potential issues with modality signals during testing (or deployment) of MSA models.

3.1 Experiment Setup

Models. In order to fully verify the universality of our experiments, we select a series of diverse SOTA models, ranging from RNN-based to Transformer-based architectures. These models work across different granularities from word-level to sentence-level variants:

(i) **MISA** (Hazarika et al., 2020) is a popular model that generates modality-invariant and -specific features of multimodal data, to learn both shared and unique characteristics of each modality. (ii) **BBFN** (Han et al., 2021a) in a similar vein performs fusion and separation to increase cross-modal relevances and differences. This work acknowledges the dominance of text modality in MSA and proposes two text-centric bi-modal transformers to increase performance. (iii) **Self-MM** (Yu et al., 2021) focuses on the relationship between multi- and uni-modal predictions by multi-tasking consistencies and differences between them. (iv) **MMIM** (Han et al., 2021b) incorporates mutual information (MI) into MSA by maximizing MI at the input and fusion level. (v) **MulT** (Tsai et al., 2019) merges multimodal time series through multiple sets of directional pairwise cross-modal transformers. It accounts for long-range dependencies across modality elements to create a strong baseline (see Appendix B).

Datasets. We consider two benchmark datasets widely used in the field of multimodal sentiment analysis, CMU-MOSI (Zadeh et al., 2016), which is a popular dataset for studying the intensity of multimodal sentiment in the MSA field and CMU-MOSEI (Bagher Zadeh et al., 2018) which is a larger counterpart of MOSI with richer annotations and more diverse samples. Both these datasets contain short utterance videos and provide language, audio, and visual modality features.

3.2 Proposed Diagnostic Checks

We propose two diagnostic checks that introduce i) *Missing modalities*, which drops (or nullifies) a modality from the input and ii) *Noisy Modalities*

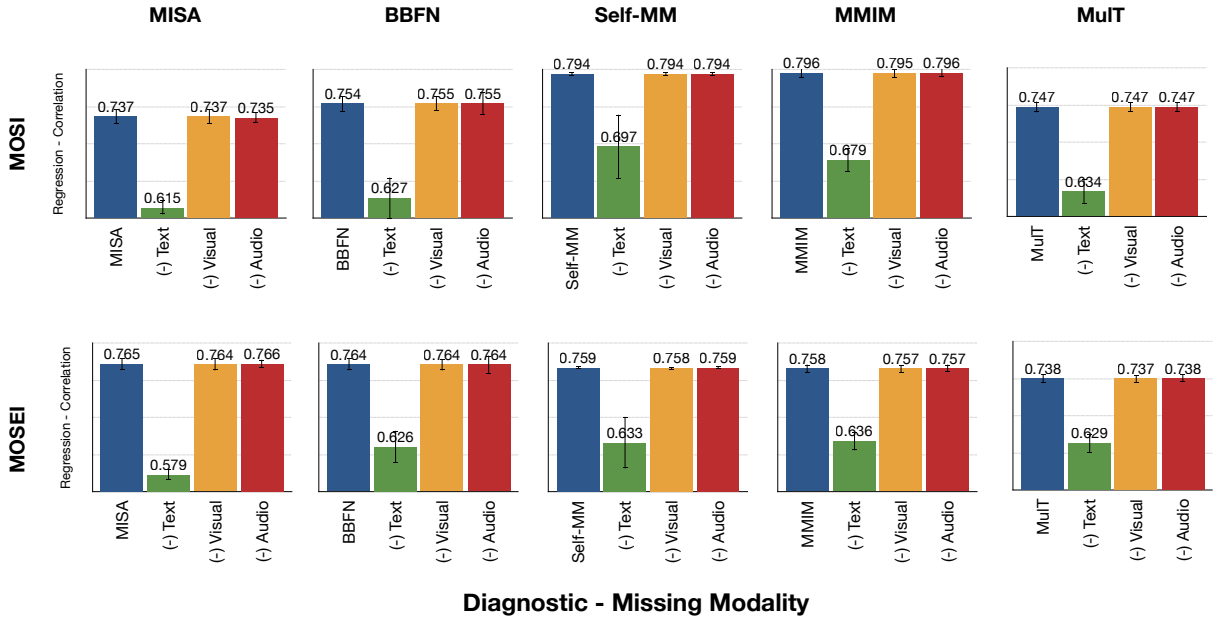

Figure 2: Diagnostic checks (missing modality) for modality robustness in MOSI and MOSEI datasets. Results are averaged over three independent runs. Each modality error is applied to 30% of testing data. For noise diagnostic, we provide the results in Fig. 3.

which include random changes to the modality representations, introduced via white Gaussian noise to the respective modality representations[1]. To simulate a realistic scenario, we apply these checks to 30% of the testing data set[2]. Given the increased dependence on language modality in MSA models, we limit our study to errors introduced only in language modality without loss of generality.

Procedure. We aim to intervene on modality representations to simulate modality errors. For the language modality l, all models map the sequence of tokens $U_l \in R^{T_l}$ to its low-level embedding $\mathbf{U}_l \in R^{T_l \times d_l}$ with T_l tokens and d_l embedding dimension. This low-level sequence is then encoded into hidden representations using an encoder of choice, such as BERT (Devlin et al., 2019), to achieve the language representation vector $\mathbf{u}_l = enc_{\theta_l}(U_l) \in \mathbb{R}^d$. We intervene on this representation and apply our diagnostics as follows.

We sample 30% of u_l from the testing set and modify them as $\hat{\mathbf{u}}_l = f(\mathbf{u}_l)$, where $f(\mathbf{x})$ is defined as either $f(\mathbf{x}) = \mathbf{x} \odot \mathbf{0}$ for modality dropping (nulling the vector to 0s by elementwise multiplication) or $f(\mathbf{x}) = \mathbf{x} + \mathcal{N}(\mathbf{0}, \mathbf{1})$ to add white noise. The modified $\hat{u}_l$ is then fed to the rest of the network as usual.

In the selected models, we apply diagnostics at different network locations. These include the representations before the hidden projection, such as in MISA, or fusion operation, such as in Self-MM. For MulT and BBFN, we apply the interventions right after the word embeddings. Detailed discussion on the location of interventions is provided in Appendix A.

Observations. Fig. 2 presents the results, where across both MOSI and MOSEI datasets, we find that language modality is highly sensitive to modality errors in the language source (across all models). This trend is observed for both missing and noisy modality checks, thus highlighting the concerns over robustness of these SOTA models. These diagnostic checks are easy to analyze, and we hope they will become an integrated part of the model-development pipeline in MSA.

4 Robust Training

In this section, we explore how to reduce the sensitivity of the models to the dominant modality, i.e., language. One of the popular ways to alleviate such issues is to *teach* the model such scenarios during training. We dub this approach as *modality-perturbation*, which is conceptually similar to removing modalities in (Ma et al., 2021) or adding noise in (Miyato et al., 2018). It sim-

[1]While missing and noisy errors are predominant in the wild, we leave other potential forms of errors, such as affine transformations to the representations for future work.

[2]We set 30% arbitrarily to simulate modality errors to a proportion of the input signals.

ulates the modality errors during training so that the model learns to expect duch events during testing/deployment. The procedure is as follows,

1. **Training:**
   (a) For a particular batch of data, sample a proportion of the data to be perturbed.
   (b) Similar to the diagnostic checks in § 3, perturb the dominant modality (in our case, language) of half of this data with *missing* and the other half with *noisy* perturbation. Repeat both these steps for the next batch.
2. **Testing:** Apply the diagnostic checks as in § 3.

This simple approach can be interpreted as regularization akin to dropouts or noising strategies used in de-noising auto-encoders.

4.1 Results

Robustness. Table 1 presents the results, where we perform *balanced* perturbation between missing and noisy modalities. For the 30% perturbable data in training, we drop the language modality on 15% and for the other 15%, we add noise. This setting improves the diagnostics in both kinds of errors. Appendix C presents results on other proportions of the training data.

With balanced perturbation, (BBFN-MOSI) reduces the relative drop on *missing* language diagnostic by 31% (in F1) and 98% on *noise*. Also, *missing* drop reduces by 11% in Corr and by 99% for *noisy* diagnostic. (Self-MM, MOSI) increases the relative drop in Corr slightly on *missing* diagnostic, but in all other cases, it is significantly reduced. For example, the F1 drop on MOSI for *noisy* diagnostic reduces significantly by 93%. Table 1 also shows that our method performs well on both RNN-based and Transformer-based models, demonstrating the wide applicability of our method.

Performance Trade-off. While alleviating robustness via regularization is well-known in the literature, there is often a trade-off with absolute performance in the original testing setup. Most approaches that achieve robustness take a hit at their best performance on clean input (Zhang et al., 2022) (Nakkiran, 2019) (Su et al., 2018) (Tsipras et al., 2019). This raises the question of whether introducing *modality-perturbation* reduces the performance of the model on the original testing set.

We find the answer to this is *No*. Surprisingly, our robust training procedure *does not* degrade in its original performance and can perform similar to the original model variants. This is highly ideal as we achieve robustness without compromising on performance in clean data.

	Diagnostic (30%)	Robust Training	MOSI		MOSEI	
			Corr	F1	Corr	F1
MISA		-	0.737	82.40	0.765	85.76
		Yes	0.736	81.42	0.767	85.97
	missing	-	↓ 0.122	↓ 11.53	↓ 0.186	↓ 8.45
		Yes	↓ 0.210	↓ 9.96	↓ 0.147	↓ 8.26
	noise	-	↓ 0.122	↓ 11.67	↓ 0.136	↓ 8.36
		Yes	↓ 0.163	↓ 10.10	↓ 0.002	↓ 0.19
BBFN		-	0.754	83.12	0.764	85.70
		Yes	0.754	83.28	0.763	85.43
	missing	-	↓ 0.127	↓ 10.55	↓ 0.139	↓ 10.57
		Yes	↓ 0.119	↓ 7.28	↓ 0.124	↓ 7.88
	noise	-	↓ 0.232	↓ 8.58	↓ 0.308	↓ 9.62
		Yes	↓ 0.046	↓ 0.16	↓ 0.003	↓ 0.23
Self-MM		-	0.794	85.61	0.759	84.62
		Yes	0.790	84.73	0.754	84.67
	missing	-	↓ 0.099	↓ 11.74	↓ 0.126	↓ 9.04
		Yes	↓ 0.120	↓ 9.66	↓ 0.122	↓ 6.86
	noise	-	↓ 0.154	↓ 8.35	↓ 0.172	↓ 9.48
		Yes	↓ 0.041	↓ 0.58	↓ 0.051	↓ 1.02
MMIM		-	0.796	86.02	0.758	84.89
		Yes	0.784	84.67	0.751	83.15
	missing	-	↓ 0.117	↓ 9.37	↓ 0.122	↓ 8.15
		Yes	↓ 0.146	↓ 10.48	↓ 0.115	↓ 6.62
	noise	-	↓ 0.197	↓ 9.55	↓ 0.191	↓ 9.18
		Yes	↓ 0.180	↓ 8.88	↓ 0.096	↓ 4.41
Mult		-	0.747	82.25	0.738	83.37
		Yes	0.744	82.21	0.745	83.95
	missing	-	↓ 0.113	↓ 12.17	↓ 0.109	↓ 6.60
		Yes	↓ 0.117	↓ 9.63	↓ 0.113	↓ 7.03
	noise	-	↓ 0.295	↓ 8.99	↓ 0.263	↓ 7.95
		Yes	↓ 0.068	↓ 6.13	↓ 0.001	↑ 0.02

Table 1: Robust Training is performed with 15% missing and 15% noise perturbation. Results are averaged over 3 random runs. More perturbation are provided in Appendix C. Higher drops between Non-robust and robust training (consecutive rows) are highlighted.

5 Conclusion

In this work, we performed a systematic study that demonstrate the double-edged nature of dominant modality in SOTA MSA models. Our analysis using diagnostic checks reveal high susceptibility to performance drops when presented with unwanted errors in their representations.

To alleviate the issues, we also study robust training methods that uses modality perturbations. Critically, we find that robustness and performance can co-exist without an explicit trade-off. These improvements demonstrate a positive nudge in the effort to achieve robustness and we believe there remains significant room for improvement. With this work, by proposing simple and easy-to-integrate diagnostic checks and training methods, we hope to permeate discussions on robustness into mainstream MSA research.

References

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 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)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. 2021. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9847–9857.

Itai Gat, Idan Schwartz, Alexander G. Schwing, and Tamir Hazan. 2020. Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Wei Han, Hui Chen, Alexander F. Gelbukh, Amir Zadeh, Louis-Philippe Morency, and Soujanya Poria. 2021a. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In *ICMI '21: International Conference on Multimodal Interaction, Montréal, QC, Canada, October 18-22, 2021*, pages 6–15. ACM.

Wei Han, Hui Chen, and Soujanya Poria. 2021b. 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, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 9180–9192. Association for Computational Linguistics.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. MISA: modality-invariant and -specific representations for multimodal sentiment analysis. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 1122–1131. ACM.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. 2021. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2302–2310.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993.

Preetum Nakkiran. 2019. Adversarial robustness may be at odds with simplicity. *CoRR*, abs/1901.00532.

Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. 2018. Is robustness the cost of accuracy? - A comprehensive study on the robustness of 18 deep image classification models. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XII*, volume 11216 of *Lecture Notes in Computer Science*, pages 644–661. Springer.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6558–6569. Association for Computational Linguistics.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2018. Learning factorized multimodal representations. In *International Conference on Learning Representations*.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. 2019. Robustness may be at odds with accuracy. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Yang Wu, Zijie Lin, Yanyan Zhao, Bing Qin, and Li-Nan Zhu. 2021. A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4730–4738, Online. Association for Computational Linguistics.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 10790–10797. AAAI Press.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88.

Ying Zeng, Sijie Mai, and Haifeng Hu. 2021. Which is making the contribution: Modulating unimodal and cross-modal dynamics for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1262–1274.

Yuxin Zhang, Jindong Wang, Yiqiang Chen, Han Yu, and Tao Qin. 2022. Adaptive memory networks with self-supervised learning for unsupervised anomaly detection. *CoRR*, abs/2201.00464.

A Model Details

MISA: We get the MISA model from its official repository [3]. In this model, we apply the interventions at the following encoded language representation from the original paper:

$$\mathbf{u}_l = \text{Bert}\left(\mathbf{U}_l; \theta_l^{bert}\right) \quad (1)$$

$$\hat{\mathbf{u}}_l = f(\hat{\mathbf{u}}_l) \quad (2)$$

$$\mathbf{h}_l^c = E_c\left(\hat{\mathbf{u}}_l; \theta^c\right), \quad \mathbf{h}_l^p = E_p\left(\hat{\mathbf{u}}_l; \theta_l^p\right) \quad (3)$$

That is, the interventions are applied before the language representation is projected to its shared and private subspaces.

BBFN: We get the BBFN model from its official repository [4]. In this model, we execute the interventions after the following language embedding from the original paper.

$$\mathbf{M}_l = (m_0, m_1, \ldots, m_{n+1}) \quad (4)$$

$$\hat{\mathbf{M}}_l = f(\mathbf{M}_l) \quad (5)$$

Self-MM: We get the Self-MM model from its official repository[5]. In this model, we set the interventions after the language features encoded below from the original paper.

$$\mathbf{F}_l = BERT\left(I_l; \theta_l^{bert}\right) \in R^{d_l} \quad (6)$$

$$\hat{\mathbf{F}}_l = f(\mathbf{F}_l) \quad (7)$$

[3]https://github.com/declare-lab/MISA/tree/ec42faddde0d210cf7368aebf2118fe9570e7102

[4]https://github.com/declare-lab/BBFN/tree/be15f947ed7539b3c54381e453f09439466ed915

[5]https://github.com/thuiar/Self-MM

Models	Item	CMU-MOSI	CMU-MOSEI
MISA	Learning rate	1e-5	4e-5
	Optimizer	RMSprop	Adam
	Activation	hardtanh	relu
BBFN	Learning rate	1e-4	5e-5
MMIM	Batch size	32	64
	learning rate η_{lld}	4e-3	1e-3
	learning rate η_{main}	1e-3	5e-4
	α	0.3	0.1
	β	0.1	0.05
	V-LSTM hidden dim	32	64
	A-LSTM hidden dim	32	16
MulT	Batch size	128	16
	Learning rate	1e-4	1e-4
	Optimizer	Adam	Adam
	Transformers Hidden Unit Size d	30	30
	No. of Crossmodal Attention Heads	10	10
	No. of Crossmodal a Blocks D	4	4
	Textual Embedding Dropout	0.3	0.3
	Crossmodal Attention Block Dropout	0.2	0.1
	Output Dropout	0.2	0.1
	Gradient Clip	0.8	1.0
	No. of Epochs	100	20
	Use Bert	Yes	Yes

Table 2: Hyper-parameter config used to train the models.

MMIM: We get the MMIM model from its official repository [6]. For this model, we perform the interventions after the following encoded language representation from the original paper.

$$\mathbf{x}_l = \text{BERT}\left(X_l; \theta_l^{BERT}\right) \quad (8)$$

$$\hat{\mathbf{x}}_l = f(\mathbf{x}_l) \quad (9)$$

Mult: We get the Mult model from its official repository [7]. We intervene in this model after the following encoded language representation:.

$$\mathbf{x}_l = \text{Conv1D}(X_l, k_l) \in \mathbb{R}^{T_l \times d} \quad (10)$$

$$\hat{\mathbf{x}}_l = f(\mathbf{x}_l) \quad (11)$$

B Reproducing Results

For each model we train the models to achieve performances close to reported in the respective papers. Table 2 presents the hyper-parameters we used to reproduce their results.

[6]https://github.com/declare-lab/Multimodal-Infomax

[7]https://github.com/yaohungt/Multimodal-Transformer

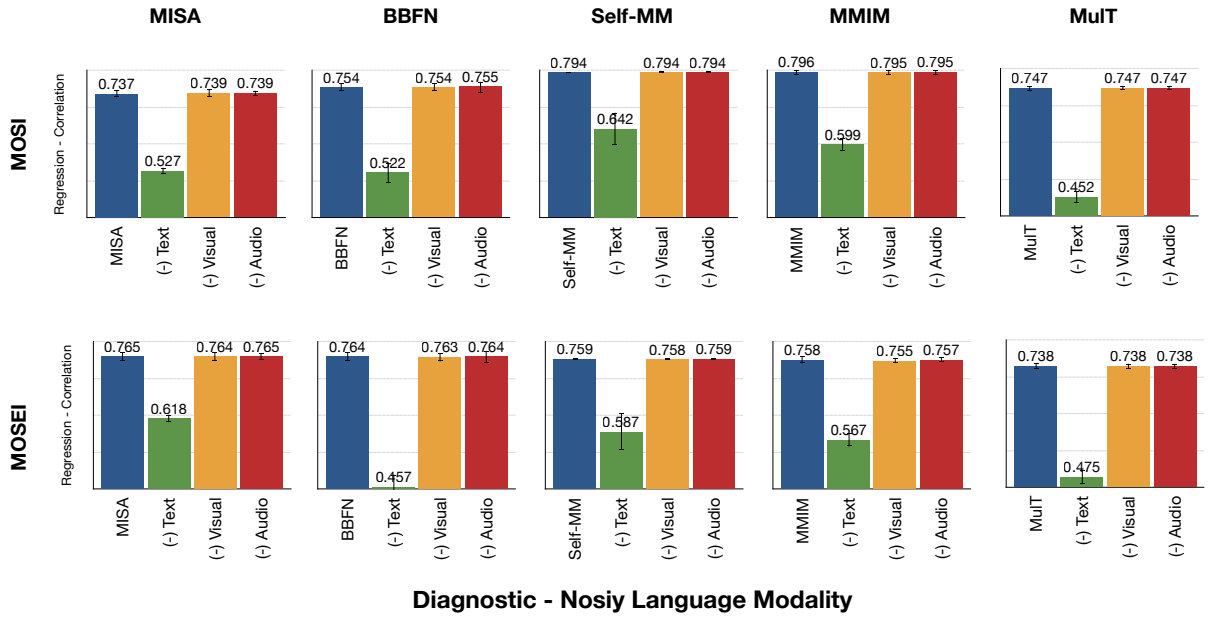

Figure 3: Diagnostic checks (noisy modality) for modality robustness in MOSI and MOSEI datasets. Results are averaged over three independent runs. Each modality error is applied to 30% of testing data.

C Additional Results on Modality-Perturbation

We also analyze with varying proportions of perturbations in the training and testing phase, respectively. As seen in Table 3, as the noise gradually increases from 5% to 15%, the drop of Corr in MOSI is gradually reduced , which shows the robustness is getting better, until it reaches the optimum at 30% perturbation (15% missing + 15% noise). In other models, 30% perturbation is also advantageous. For example, in Table 7, (Mult, MOSEI) reduces Corr drop while improving F1 performance in 30% perturbation. Although it is only a small improvement at present, we believe that there will be more meaningful improvements in the future.

	Robust Training	Diagnostic	MOSI		MOSEI	
			Corr	F1	Corr	F1
MISA	-		0.737	82.40	0.765	85.76
		missing 5%	↓ 0.011	↓ 1.18	↓ 0.042	↓ 1.57
		noise 5%	↓ 0.022	↓ 0.75	→ 0	→ 0
		missing 10%	↓ 0.031	↓ 2.19	↓ 0.073	↓ 2.41
		noise 10%	↓ 0.047	↓ 2.56	↓ 0.031	↓ 1.75
		missing 15%	↓ 0.050	↓ 3.99	↓ 0.126	↓ 4.10
		noise 15%	↓ 0.076	↓ 4.05	↓ 0.068	↓ 3.88
		missing 30%	↓ 0.122	↓ 11.53	↓ 0.186	↓ 8.45
		noise 30%	↓ 0.210	↓ 9.96	↓ 0.147	↓ 8.26
	5% missing 5% noise		0.714	80.99	0.765	85.68
		missing 5%	↓ 0.010	↓ 1.38	↓ 0.042	↓ 1.57
		noise 5%	↓ 0.032	↓ 1.37	→ 0	→ 0
		missing 10%	↓ 0.030	↓ 2.17	↓ 0.074	↓ 2.41
		noise 10%	↓ 0.055	↓ 2.44	↓ 0.041	↓ 2.22
		missing 15%	↓ 0.044	↓ 4.52	↓ 0.126	↓ 0.41
		noise 15%	↓ 0.094	↓ 3.65	↓ 0.058	↓ 3.53
		missing 30%	↓ 0.121	↓ 11.08	↓ 0.186	↓ 8.37
		noise 30%	↓ 0.201	↓ 8.71	↓ 0.150	↓ 8.12
	10% missing 10% noise		0.734	81.40	0.765	85.68
		missing 5%	↓ 0.011	↓ 1.04	↓ 0.022	↓ 1.58
		noise 5%	↓ 0.028	↓ 1.20	→ 0	→ 0
		missing 10%	↓ 0.030	↓ 2.05	↓ 0.042	↓ 2.54
		noise 10%	↓ 0.045	↓ 2.27	↓ 0.001	↓ 0.02
		missing 15%	↓ 0.049	↓ 3.54	↓ 0.071	↓ 4.04
		noise 15%	↓ 0.073	↓ 3.42	↓ 0.001	↓ 0.43
		missing 35%	↓ 0.118	↓ 11.11	↓ 0.135	↓ 8.17
		noise 30%	↓ 0.181	↓ 9.55	↓ 0.035	↓ 0.54
	15% missing 15% noise		0.736	81.42	0.767	85.97
		missing 5%	↓ 0.012	↓ 1.50	↓ 0.021	↓ 1.46
		noise 5%	↓ 0.027	↓ 1.50	→ 0	→ 0
		missing 10%	↓ 0.033	↓ 2.71	↓ 0.041	↓ 2.43
		noise 10%	↓ 0.032	↓ 2.12	↓ 0.001	↓ 0.03
		missing 15%	↓ 0.052	↓ 4.40	↓ 0.021	↓ 4.11
		noise 15%	↓ 0.073	↓ 4.99	↓ 0.001	↓ 0.12
		missing 30%	↓ 0.122	↓ 11.67	↓ 0.136	↓ 8.36
		noise 30%	↓ 0.163	↓ 10.10	↓ 0.002	↓ 0.19

Table 3: MISA Robust Training. Results are averaged over three random runs.

	Robust Training	Diagnostic	MOSI		MOSEI	
			Corr	F1	Corr	F1
BBFN	-		0.754	83.12	0.764	85.70
		missing 5%	↓ 0.013	↓ 1.51	↓ 0.020	↓ 1.97
		noise 5%	↓ 0.034	↓ 0.75	↓ 0.050	↓ 1.21
		missing 10%	↓ 0.028	↓ 2.59	↓ 0.038	↓ 3.34
		noise 10%	↓ 0.093	↓ 2.90	↓ 0.121	↓ 3.54
		missing 15%	↓ 0.032	↓ 3.79	↓ 0.055	↓ 5.45
		noise 15%	↓ 0.080	↓ 2.28	↓ 0.154	↓ 4.18
		missing 30%	↓ 0.127	↓ 10.55	↓ 0.139	↓ 10.57
		noise 30%	↓ 0.232	↓ 8.58	↓ 0.308	↓ 9.62
	5% missing 5% noise		0.743	82.39	0.765	85.48
		missing 5%	↓ 0.020	↓ 0.94	↓ 0.017	↓ 1.00
		noise 5%	↓ 0.002	→ 0	→ 0	↓ 0.08
		missing 10%	↓ 0.039	↓ 2.17	↓ 0.032	↓ 1.95
		noise 10%	→ 0	→ 0	→ 0	↓ 0.06
		missing 15%	↓ 0.045	↓ 2.18	↓ 0.050	↓ 3.57
		noise 15%	↓ 0.002	↓ 0.15	↓ 0.001	→ 0
		missing 30%	↓ 0.049	↓ 1.92	↓ 0.122	↓ 7.60
		noise 30%	→ 0	↓ 0.20	↓ 0.001	↓ 0.04
	10% missing 10% noise		0.742	81.66	0752	85.15
		missing 5%	↓ 0.018	↓ 1.66	↓ 0.018	↓ 1.62
		noise 5%	↓ 0.001	→ 0	↓ 0.022	↓ 0.05
		missing 10%	↓ 0.034	↓ 2.87	↓ 0.035	↓ 2.87
		noise 10%	↓ 0.001	→ 0	↓ 0.003	↓ 0.09
		missing 15%	↓ 0.036	↓ 3.63	↓ 0.050	↓ 4.47
		noise 15%	↓ 0.001	↓ 0.01	↓ 0.004	↓ 0.07
		missing 30%	↓ 0.126	↓ 9.75	↓ 0.125	↓ 9.40
		noise 30%	↓ 0.003	↓ 0.46	↓ 0.079	↓ 0.31
	15% missing 15% noise		0.754	83.28	0.763	85.43
		missing 5%	↓ 0.020	↓ 1.50	↓ 0.018	↓ 0.96
		noise 5%	↓ 0.001	→ 0	↓ 0.001	→ 0
		missing 10%	↓ 0.031	↓ 2.26	↓ 0.036	↓ 2.08
		noise 10%	↓ 0.002	→ 0	→ 0	↓ 0.05
		missing 15%	↓ 0.035	↓ 2.71	↓ 0.057	↓ 3.62
		noise 15%	↓ 0.003	↓ 0.14	↓ 0.001	→ 0
		missing 30%	↓ 0.119	↓ 7.28	↓ 0.124	↓ 7.88
		noise 30%	↓ 0.046	↓ 0.16	↓ 0.003	↓ 0.23

Table 4: BBFN Robust Training. Results are averaged over three random runs.

	Robust Training	Diagnostic	MOSI		MOSEI	
			Corr	F1	Corr	F1
Self-MM	-		0.794	85.61	0.759	84.62
		missing 5%	↓ 0.023	↓ 1.93	↓ 0.018	↓ 1.99
		noise 5%	↓ 0.009	↓ 0.91	↓ 0.022	↓ 1.32
		missing 10%	↓ 0.046	↓ 3.57	↓ 0.040	↓ 3.51
		noise 10%	↓ 0.039	↓ 2.43	↓ 0.058	↓ 2.91
		missing 15%	↓ 0.051	↓ 4.77	↓ 0.056	↓ 4.51
		noise 15%	↓ 0.050	↓ 2.71	↓ 0.069	↓ 3.65
		missing 30%	↓ 0.099	↓ 11.74	↓ 0.126	↓ 9.04
		noise 30%	↓ 0.154	↓ 8.35	↓ 0.172	↓ 9.48
	5% missing 5% noise		0.798	83.97	0.789	0.837
		missing 5%	↓ 0.022	↓ 1.73	↓ 0.021	↓ 1.83
		noise 5%	↓ 0.002	→ 0	↓ 0.002	↓ 0.05
		missing 10%	↓ 0.046	↓ 3.32	↓ 0.046	↓ 3.28
		noise 10%	↓ 0.012	↓ 0.14	↓ 0.011	↓ 0.44
		missing 15%	↓ 0.050	↓ 4.63	↓ 0.053	↓ 4.60
		noise 15%	↓ 0.018	↓ 0.09	↓ 0.018	↓ 0.51
		missing 30%	↓ 0.119	↓ 9.57	↓ 0.124	↓ 9.77
		noise 30%	↓ 0.046	↓ 0.33	↓ 0.043	↓ 1.04
	10% missing 10% noise		0.789	83.67	0.764	0.849
		missing 5%	↓ 0.021	↓ 1.83	↓ 0.017	↓ 0.63
		noise 5%	↓ 0.002	↓ 0.05	↓ 0.005	↓ 0.03
		missing 10%	↓ 0.046	↓ 3.28	↓ 0.038	↓ 1.93
		noise 10%	↓ 0.011	↓ 0.44	↓ 0.011	↓ 0.15
		missing 15%	↓ 0.053	↓ 4.60	↓ 0.057	↓ 3.16
		noise 15%	↓ 0.018	↓ 0.51	↓ 0.019	↓ 0.13
		missing 30%	↓ 0.124	↓ 9.77	↓ 0.126	↓ 7.42
		noise 30%	↓ 0.043	↓ 1.04	↓ 0.056	↓ 1.01
	15% missing 15% noise		0.790	84.73	0.754	84.67
		missing 5%	↓ 0.022	↓ 1.91	↓ 0.017	↓ 0.60
		noise 5%	↓ 0.003	→ 0	↓ 0.004	↓ 0.01
		missing 10%	↓ 0.046	↓ 3.39	↓ 0.038	↓ 1.91
		noise 10%	↓ 0.010	↓ 0.44	↓ 0.011	↓ 0.14
		missing 15%	↓ 0.049	↓ 4.58	↓ 0.055	↓ 3.08
		noise 15%	↓ 0.018	↓ 0.28	↓ 0.016	↓ 0.16
		missing 30%	↓ 0.120	↓ 9.66	↓ 0.122	↓ 6.86
		noise 30%	↓ 0.041	↓ 0.58	↓ 0.051	↓ 1.02

Table 5: Self-MM Robust Training. Results are averaged over three random runs.

	Robust Training	Diagnostic	MOSI		MOSEI	
			Corr	F1	Corr	F1
MMIM	-		0.796	86.02	0.758	84.89
		missing 5%	↓ 0.028	↓ 1.34	↓ 0.016	↓ 0.93
		noise 5%	↓ 0.035	↓ 1.52	↓ 0.031	↓ 1.71
		missing 10%	↓ 0.067	↓ 4.16	↓ 0.034	↓ 2.43
		noise 10%	↓ 0.058	↓ 2.10	↓ 0.070	↓ 2.64
		missing 15%	↓ 0.056	↓ 2.78	↓ 0.056	↓ 4.13
		noise 15%	↓ 0.078	↓ 4.65	↓ 0.094	↓ 4.67
		missing 30%	↓ 0.117	↓ 9.37	↓ 0.122	↓ 8.15
		noise 30%	↓ 0.197	↓ 9.55	↓ 0.191	↓ 9.18
	5% missing 5% noise		0.797	85.13	0.755	0.836
		missing 5%	↓ 0.057	↓ 1.19	↓ 0.016	↓ 0.62
		noise 5%	↓ 0.021	↓ 1.20	↓ 0.025	↓ 1.07
		missing 10%	↓ 0.035	↓ 1.80	↓ 0.046	↓ 2.29
		noise 10%	↓ 0.075	↓ 2.25	↓ 0.056	↓ 2.43
		missing 15%	↓ 0.057	↓ 4.37	↓ 0.168	↓ 2.26
		noise 15%	↓ 0.072	↓ 3.96	↓ 0.063	↓ 2.75
		missing 30%	↓ 0.117	↓ 8.24	↓ 0.120	↓ 6.52
		noise 30%	↓ 18.75	↓ 9.31	↓ 0.178	↓ 7.34
	10% missing 10% noise		0.794	84.76	0.758	84.81
		missing 5%	↓ 0.020	↓ 1.50	↓ 0.016	↓ 1.07
		noise 5%	↓ 0.006	↓ 0.46	↓ 0.024	↓ 1.28
		missing 10%	↓ 0.032	↓ 2.09	↓ 0.043	↓ 2.79
		noise 10%	↓ 0.060	↓ 2.57	↓ 0.045	↓ 2.58
		missing 15%	↓ 0.060	↓ 3.27	↓ 0.058	↓ 3.84
		noise 15%	↓ 0.062	↓ 3.65	↓ 0.063	↓ 3.42
		missing 30%	↓ 0.110	↓ 8.40	↓ 0.119	↓ 7.50
		noise 30%	↓ 0.163	↓ 9.18	↓ 0.177	↓ 8.16
	15% missing 15% noise		0.784	84.67	0.751	83.15
		missing 5%	↓ 0.014	↓ 1.79	↓ 0.020	↓ 0.89
		noise 5%	↓ 0.028	↓ 1.04	↓ 0.016	↓ 0.69
		missing 10%	↓ 0.038	↓ 2.18	↓ 0.030	↓ 2.24
		noise 10%	↓ 0.050	↓ 1.30	↓ 0.028	↓ 0.62
		missing 15%	↓ 0.061	↓ 4.91	↓ 0.059	↓ 2.77
		noise 15%	↓ 0.071	↓ 4.08	↓ 0.040	↓ 1.58
		missing 30%	↓ 0.146	↓ 10.48	↓ 0.115	↓ 6.62
		noise 30%	↓ 0.180	↓ 8.88	↓ 0.096	↓ 4.41

Table 6: MMIM Robust Training. Results are averaged over three random runs.

	Robust Training	Diagnostic	MOSI		MOSEI	
			Corr	F1	Corr	F1
MulT	-		0.747	82.25	0.738	83.37
		missing 5%	↓ 0.032	↓ 3.76	↓ 0.019	↓ 0.84
		noise 5%	↓ 0.053	↓ 1.67	→ 0	→ 0
		missing 10%	↓ 0.046	↓ 3.92	↓ 0.031	↓ 1.85
		noise 10%	↓ 0.069	↓ 3.03	→ 0	→ 0
		missing 15%	↓ 0.052	↓ 4.84	↓ 0.051	↓ 3.29
		noise 15%	↓ 0.172	↓ 5.00	↓ 0.152	↓ 4.50
		missing 30%	↓ 0.113	↓ 12.17	↓ 0.109	↓ 6.60
		noise 30%	↓ 0.295	↓ 8.99	↓ 0.263	↓ 7.95
	5% missing 5% noise		0.748	81.90	0.748	84.59
		missing 5%	↓ 0.031	↓ 1.50	↓ 0.018	↓ 0.93
		noise 5%	↓ 0.021	↓ 1.67	→ 0	→ 0
		missing 10%	↓ 0.043	↓ 2.43	↓ 0.030	↓ 1.99
		noise 10%	↓ 0.021	↓ 1.96	→ 0	→ 0
		missing 15%	↓ 0.050	↓ 3.36	↓ 0.050	↓ 3.51
		noise 15%	↓ 0.067	↓ 3.30	↓ 0.173	↓ 3.97
		missing 30%	↓ 0.077	↓ 5.12	↓ 0.074	↓ 4.90
		noise 30%	↓ 0.094	↓ 5.17	↓ 0.198	↓ 5.50
	10% missing 10% noise		0.741	81.31	0.746	84.13
		missing 5%	↓ 0.034	↓ 2.57	↓ 0.019	↓ 0.96
		noise 5%	↓ 0.012	↓ 0.09	→ 0	→ 0
		missing 10%	↓ 0.046	↓ 3.06	↓ 0.030	↓ 0.019
		noise 10%	↓ 0.031	↓ 2.42	→ 0	→ 0
		missing 15%	↓ 0.053	↓ 5.52	↓ 0.048	↓ 3.19
		noise 15%	↓ 0.045	↓ 4.08	↓ 0.153	↓ 4.08
		missing 30%	↓ 0.077	↓ 6.57	↓ 0.072	↓ 4.62
		noise 30%	↓ 0.056	↓ 4.57	↓ 0.202	↓ 5.58
	15% missing 15% noise		0.744	82.21	0.745	83.95
		missing 5%	↓ 0.034	↓ 2.56	↓ 0.018	↓ 0.86
		noise 5%	↓ 0.023	↓ 1.52	→ 0	→ 0
		missing 10%	↓ 0.047	↓ 3.93	↓ 0.032	↓ 1.86
		noise 10%	↓ 0.039	↓ 3.62	→ 0	→ 0
		missing 15%	↓ 0.052	↓ 5.15	↓ 0.051	↓ 3.27
		noise 15%	↓ 0.046	↓ 3.62	→ 0	↓ 0.08
		missing 30%	↓ 0.117	↓ 9.63	↓ 0.113	↓ 7.03
		noise 30%	↓ 0.068	↓ 6.13	↓ 0.001	↑ 0.02

Table 7: MulT Robust Training. Results are averaged over three random runs.