MODALITY COMPLEMENTARITY: TOWARDS UNDERSTANDING MULTI-MODAL ROBUSTNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Along with the success of multi-modal learning, the robustness of multi-modal learning is receiving attention due to real-world safety concerns. Multi-modal models are anticipated to be more robust due to the possible redundancy between modalities. However, some empirical results have offered contradictory conclusions. In this paper, we point out an essential factor that causes this discrepancy: The difference in the amount of modality-wise complementary information. We provide an information-theoretical analysis of how the modality complementarity affects the multi-modal robustness. Based on the analysis, we design a metric for quantifying how complementary the modalities are to others and propose an effective pipeline to calculate our metric. Experiments on carefully-designed synthetic data verify our theory. Further, we apply our metric to real-world multi-modal datasets and reveal their property. To our best knowledge, we are the first to identify modality complementarity as an important factor affecting multi-modal robustness.

1 INTRODUCTION

Recently, deep neural networks have proved successful in various areas, such as image recognition (He et al., 2015; Krizhevsky et al., 2012), speech recognition (Chorowski et al., 2015) and neural machine translation (Wu et al., 2016). The revolution is also happening in multi-modal research, *e.g.* RGB-D semantic segmentation (Wang et al., 2016), audio-visual learning (Zhao et al., 2018), and visual question answering (Antol et al., 2015). Intuitively, multi-modal models are anticipated to be more robust due to the potential redundancy between modalities. When one of the modalities is corrupted, others can compensate for the loss. This intuition is supported by both psychological studies of the human perception system (Sumby & Pollack, 1954) and deep learning practices (Zhang et al., 2019b; Qian et al., 2021; Wang et al., 2020).

However, some recent studies cast doubt on this belief. From a theoretical perspective, the multi-modal models usually have a larger input dimension than uni-modal models, and the increase of input dimensions significantly degrades model robustness (Ford et al., 2019; Simon-Gabriel et al., 2019). From an empirical view, some experiments suggest that multi-modal integration may be more vulnerable to attacks or corruptions than uni-modal models (Yu et al., 2020; Tian & Xu, 2021; Ma et al., 2022).

What causes this contradiction in multi-modal robustness? We notice that the conclusions above are drawn under assorted multi-modal task settings ranging from action classification to question answering, which vary in the presence and type of modality interconnections (Liang et al., 2022). Therefore, a question arises naturally:

What aspects of modality interconnection affect the multi-modal robustness?

We hypothesize that the *complementarity of modalities* plays an essential role. If the complementary part of each modality is negligible, the corruption of one modality would not severely damage the model performance. Otherwise, the multi-modal model could perform even worse than a uni-modal model. For the visual question answering task, the two modalities are highly complementary: Only perceiving either the question or the image could not lead to an ideal answer (Agrawal et al., 2018). For the action classification task, the RGB and optical flow are less complementary since each of them can suggest a roughly correct answer (Feichtenhofer et al., 2016b).

To validate the above hypothesis, we first demonstrate the key role of modality complementarity to model robustness through theoretical analysis. Following previous work (Tsai et al., 2020; Sun et al., 2020; Sridharan & Kakade, 2008; Tosh et al., 2021), we use an information-theoretical framework for multi-modal learning and study how the complementary information affects robustness under missing and noisy modality settings. Based on the analysis, we design a novel metric and a practical calculation pipeline built on Mutual Information Neural Estimator (MINE) (Belghazi et al., 2018) to quantify the complementarity of modalities in multi-modal datasets.

With the specially designed metric and pipeline on hand, we verify our theory and the effectiveness of our proposed metric on synthetic data and a carefully-designed toy dataset AAV-MNIST. The results are consistent with the model robustness in modality missing, noisy modality, and adversarial attack settings on the datasets we test on. Then we apply our metric to real-world multi-modal datasets to further investigate the modality complementarity in different settings. To our best knowledge, we are the first to identify and prove the important role of modality complementarity in multi-modal robustness. Hence, for future research, we recommend that researchers consider the modality complementarity as a control variable for a fairer comparison of multi-modal robustness.

The main contributions are highlighted as follows:

- We point out the effect of modality complementarity on multi-modal model robustness through information-theoretical analysis.
- We propose a dataset-wise metric to qualitatively evaluate how complementary the modalities are in each multi-modal dataset, and also design a pipeline for computing the metric in real-world datasets.
- We create a synthetic dataset and a toy dataset (AAV-MNIST) to test our metric and pipeline. These datasets cover various complementary situations of different modalities and are used to verify the effectiveness of our pipeline.
- We further reveal the modality complementarity and its relationship with model robustness in real-world multi-modal datasets, which could lead to a less biased comparison for multi-modal robustness.

2 Related Work

Multi-modal learning. Various multi-modal learning tasks and models are proposed in recent years (Baltrusaitis et al., 2017; Liang et al., 2021), such as multi-modal reasoning (Yi et al., 2019; Johnson et al., 2016), cross-modal retrieval (Gu et al., 2017; Radford et al., 2021), and cross-modal translation (Ramesh et al., 2021). Among these settings, we mainly focus on the supervised multi-modal classification setting. The theoretical understanding of multi-modal learning is relatively under-explored, with (Huang et al., 2021) deriving generalization error bounds and (Sun et al., 2020) comparing with the Bayesian posterior classifiers. A concept close to multi-modal learning is the multi-view learning (Xu et al., 2013). The theory of multi-view learning has long been studied both theoretically (Zhang et al., 2019a; Tosh et al., 2021) and empirically (Sindhwani et al., 2005; Ding et al., 2021; Amini et al., 2009; Tian et al., 2019). Earlier work (Kakade & Foster, 2007; Sridharan & Kakade, 2008) proposes the multi-view assumption: Each modality suffices to predict the label. Recently, many multi-view analyses adopted this assumption (Han et al., 2021; Tsai et al., 2020; Lin et al., 2021; Federici et al., 2020; Lin et al., 2022). However, as pointed out by (Huang et al., 2021; 2022), this might not hold in the multi-modal learning setting.

Model robustness. Model robustness under data missing (Ramoni & Sebastiani, 2001), random corruption (Hendrycks & Dietterich, 2019), and adversarial attacks (Madry et al., 2017) is constantly been concerned in consideration of real-world safety issues. For uni-modal models, several methods are proposed to strengthen model robustness (Papernot et al., 2015; Huang et al., 2015; Meng & Chen, 2017). For multi-modal models, some papers regard the use of multi-modality as a way to improve robustness (Zhang et al., 2019b; Qian et al., 2021; Wang et al., 2020), while others continue to improve multi-modal models' robustness by designing new network architectures and fusion methods (Kim & Ghosh, 2019a; Tsai et al., 2018; Yang et al., 2021) and training routines (Eitel et al., 2015; Ma et al., 2021). When dealing with known missing patterns, researchers explore additional ways: data imputation through available modalities or views (Tran et al., 2017; Lin et al., 2021), or training different models for different availability of modalities (Yuan et al., 2012). Our analysis

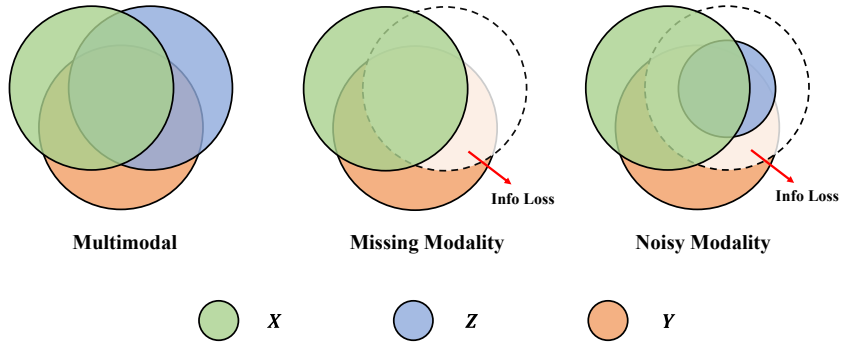

Figure 1: Illustration of relationships between the inputs and the target of a multi-modal task in different cases. X and Z are random variables representing the input of two modalities. Y is the target we would like to infer. The Info Loss refers to the loss of Y-relevant information provided by inputs, which is caused by missing or corruption of modality Z.

points out the upper bound of these methods. Apart from improving robustness, another line of work aims to analyze or estimate the robustness of existing uni-modal methods (Cohen et al., 2019; Carlini et al., 2019; Mahmood et al., 2021) and multi-modal methods (Yu et al., 2020; Tian & Xu, 2021; Ma et al., 2022; Rosenberg et al., 2021; Li et al., 2020). We in this work analyze one factor of multi-modal model robustness both theoretically and empirically.

Mutual information in deep learning. Mutual information is tightly related to deep learning through multiple ways, including information bottleneck method (Tishby et al., 2000), analysis of learning methods (Wu & Verdu, 2012; Tsai et al., 2020; Shwartz-Ziv & Tishby, 2017), and new learning methods based on mutual information (Hjelm et al., 2018; Bachman et al., 2019; Sun et al., 2020). On the other hand, learning methods help to estimate the amount of mutual information. Representative work includes Mutual Information Neural Estimator (MINE) (Belghazi et al., 2018), CPC (van den Oord et al., 2018), DIM (Hjelm et al., 2018), and DoE estimator (McAllester & Stratos, 2018). We apply the MINE to our calculation pipeline for its simplicity and effectiveness.

3 Theoretical Analysis

In this section, we first build an information-theoretical framework for multi-modal learning and show the impact of complementary information to model robustness in modality missing and single noisy source cases, which are commonly studied in previous work (Kim & Ghosh, 2019a; Tian & Xu, 2021) and widely encountered in practice, e.g., some sensors are broken, prone to noise (e.g. cameras in foggy environments), or expensive to use(e.g. X-ray data for medical analysis). An illustration of these cases is plotted in Figure 1.

3.1 Preliminaries

Notations. We use $H(A)$ to represent the entropy of a random variable A, $H(A|B)$ for the conditional entropy given another variable B, $I(A;B)$ for the mutual information between random variable A and B, $I(A;B|C)$ for the conditional mutual information conditioned on random variable C, and $I(A;B;C)$ for the interaction information (i.e., mutual information of three variables, possibly a negative value).

Multi-modal learning formulation. We adopt the formulation for multi-modal learning problem proposed in (Huang et al., 2021) Denote the M-modality input space as $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \mathcal{X}_M$ and the target space as $\mathcal{Y}$. Each data point $(X_1, X_2, \ldots, X_M, Y)$ is sampled from an unknown distribution on $\mathcal{X} \times \mathcal{Y}$. Our goal is that, based on the random input variables $X_1, X_2, \ldots, X_M$ from M modalities, we would like to infer the target Y. In classification tasks, Y is a discrete random variable, while in regression tasks Y is continuous. For instance, considering audio-visual action recognition (Gao et al., 2019; Feichtenhofer et al., 2016a), let X_1 be the audio part and X_2 be the frames of a video clip, and we want to infer the label Y, i.e., what kind of action is performed in the clip. In the subsequent analysis, we will focus on the common case $M = 2$ (Feichtenhofer et al., 2016a) for simplicity, and we denote the two modalities as $X \in \mathcal{X}$ and $Z \in \mathcal{Z}$ respectively. Notice

that our analysis and results can be extended to cases with more than two modalities at the expense of notations.

Complementary information. Now we define the *complementary information* in the following, which is essential through our theoretical analysis.

Definition 1 (complementary information)**.** *For input variables X, Z and the target Y, define the complementary information provided by X, Z as follows*

$$\Gamma_{X,Y} = I(X;Y|Z)$$
$$\Gamma_{Z,Y} = I(Z;Y|X)$$

When the target is clear from the context, we omit the Y in the subscript.

Mathematically, $I(X;Y|Z)$ represents the information in the target Y that is accessible for X but not predictable for Z. Thus Γ_X can characterize the unique label information owned by modality X, and similarly for Γ_Z. Hence Γ_X together with Γ_Z can determine the complementarity of modality X and Z. Clearly, larger Γ_X and Γ_Z imply higher complementary information content.

From the standard derivation in information theory, we can obtain the following relation:

$$I(X,Z;Y) = \Gamma_X + \Gamma_Z + I(X;Y;Z) \tag{1}$$

Previous theoretical analyses of multi-view learning Sridharan & Kakade (2008); Xu et al. (2013); Tosh et al. (2021) usually adopt the multi-view assumption that each view is redundant in terms of predicting the target, i.e. Γ_X and Γ_Z are both small. However, this assumption does not always hold in the multi-modal learning setting Antol et al. (2015). In the following subsections, we will show how the complementary information Γ_X and Γ_Z affect the model robustness in missing modality and noise settings. Motivated by this theoretical observation, we will propose a metric to evaluate the modality complementarity and a pipeline for calculation in Section 4.

Bayes error rate. We introduce the Bayes error rate Fukunaga & Hummels (1987) to measure the model performance, which is the lowest possible error for any arbitrary classifier or predictor from the multiple modalities to infer the target. Formally, given two modalities X and Z, the multi-modal Bayes errors for classification P_{e_c} and regression P_{e_r} are defined as follows:

$$P_{e_c} := \mathbb{E}_{x,z\sim P_{X,Z}}[1 - \max_{y\in Y} P(Y=y|x,z)]$$
$$P_{e_r} := \mathbb{E}_{x,z,y\sim P_{X,Z,Y}}[(y - \mathbb{E}[Y|x,z])^2]$$

The Bayes error rate helps us focus on the interconnection among modalities X, Z and target Y in each multi-modal task and omit other factors' effects on model robustness, e.g. dataset size, training routines, and network architectures.

3.2 Missing Modality

We first consider the missing modality scenario and assume modality Z is missing w.l.o.g.. Then the Bayes error rates for missing modality, denoted as $P_{e_c}^{\text{Miss}}$ and $P_{e_r}^{\text{Miss}}$ become

$$P_{e_c}^{\text{Miss}} = \mathbb{E}_{x\sim P_X}[1 - \max_{y\in Y} P(Y=y|x)]$$
$$P_{e_r}^{\text{Miss}} = \mathbb{E}_{x,y\sim P_{X,Y}}[(y - \mathbb{E}[Y|x])^2].$$

Now we establish the following theoretical guarantees to quantify differences between the Bayes error rate of multi-modal and missing-modality.

Theorem 3.1. *For random variables X, Z and discrete random variable Y, we have*

$$\frac{H(Y|X,Z) - \log 2}{\log|Y|} \le P_{e_c} \le 1 - \exp(-H(Y|X,Z)) \tag{2}$$
$$\frac{H(Y|X,Z) + \Gamma_Z - \log 2}{\log|Y|} \le P_{e_c}^{Miss} \le 1 - \exp(-H(Y|X,Z) - \Gamma_Z) \tag{3}$$

For continuous random variable Y, if we further assume that Y takes value in $[-1,1]$, then we have

$$P_{e_r}^{Miss} - P_{e_r} \le \frac{1}{2}\Gamma_Z \tag{4}$$

□

Remark 1. The **gap** between $P_{e_c}^{\text{Miss}}$ (best model performance in modality missing setting) and P_{e_c} (best model performance in normal setting) reflects the best model robustness against modality missing. For the classification task, when $\Gamma_Z = 0$, i.e., there is no complementary information of Z, the information from Z can be covered by the information from X for predicting Y. In this case, the $P_{e_c}^{\text{Miss}}$ shares the same lower and upper bound with P_{e_c}, so the performance of the best model would not be affected by modality missing. As the Γ_Z increases, the bounds for $P_{e_c}^{\text{Miss}}$ rise, while the bounds for P_{e_c} is fixed, indicating that the best model performance drops under modality missing, i.e. the robustness decays. Considering the extreme case when Γ_Z is large enough, the lower bound of $P_{e_c}^{\text{Miss}}$ is greater than the upper bound of P_{e_c}, so the missing modality performance is provably worse than normal performance.

Remark 2. For the regression task, the closer $P_{e_r}^{\text{Miss}}$ and P_{e_r} are, the robust the best model is. From the result above, the gap between two Bayes optimal predictors is bounded above by the complementary information, hence increased by Γ_Z. So the model robustness under modality missing is worsened along with the increase of Γ_Z.

3.3 Single Noisy Modality

The modality corrupted by noise is another situation that we often encounter in practice, e.g., the foggy weather results in noisy RGB images in autonomous driving. In this section, we study the case that one of the modalities has additional noise, which can be easily extended to the case that all modalities are noisy at the expense of notations. Formally, we consider that Gaussian noise N is added to the input modality Z (Zheng et al., 2016; Kim & Ghosh, 2019b). We use $R_N(Z) = Z + N$ to denote the modality Z after adding Gaussian noise. By (Cover, 1999) we can obtain the following characterization for the mutual information between Z and $R_N(Z)$

Proposition 1. *If $Z, N \in \mathbb{R}$, assuming that $0 < \mathbb{E}[Z^2] \leq p_Z, N \sim \mathcal{N}(0, \sigma)$, and N is independent of Z, then we have*

$$I(Z; R_N(Z)) \leq \frac{1}{2} \log(1 + \frac{p_Z}{\sigma}) \tag{5}$$

Remark 3. When the noise is heavy, i.e., the σ is large, the upper bound of $I(Z; R_N(Z))$ decays, indicating that it is harder to recover Z from $R_N(Z)$ and thus harder to infer Y from $R_N(Z)$. When the noise becomes very heavy, $I(Z; R_N(Z))$ will be near zero and $R_N(Z)$ is close to pure Gaussian noise, as if the modality Z is missing, which suits our intuition. In this extreme case, we can refer to the analysis in section 3.2.

In this setting, the Bayes error rate for classification denoted as $P_{e_c}^{\text{No}}$ can be written as:

$$P_{e_c}^{\text{No}} = \mathbb{E}_{x,z \sim P_{X,Z}}[1 - \max_{y \in Y} P(Y = y | x, R_N(z))]$$

Then we can provide the lower bound for $P_{e_c}^{\text{No}}$.

Theorem 3.2. *For random variables X, Y, Z, N, if $\mathbb{E}[Z^2] \leq p_Z, N \sim \mathcal{N}(0, \sigma)$, then*

$$P_{e_c}^{No} \geq \frac{H(Y|X,Z) + \Gamma_Z + I(X;Y;R_N(Z)) - \frac{1}{2}\log(4 + \frac{4p_Z}{\sigma})}{\log|Y|} \tag{6}$$

□

Remark 4. Similar to the analysis in modality missing setting, the gap between $P_{e_c}^{\text{No}}$ (best model performance in noisy modality setting) and P_{e_c} (best model performance in normal setting) reflects the best model robustness against noisy modality. For the classification task, the lower bound of $P_{e_c}^{\text{No}}$ increases as Γ_Z or σ increases. Since the bounds of P_{e_c} are fixed, the gap between $P_{e_c}^{\text{No}}$ and P_{e_c} becomes larger, and the model robustness under noisy setting is worse. Therefore, if the Γ_Z is larger, the best predictor become more vulnerable to the added noise.

4 Metric

In this section, we propose a dataset-wise metric based on the complementary information to quantify the modality complementarity. We also bring our metric to practical use by leveraging the existing mutual information estimator, Mutual Information Neural Estimator (MINE) (Belghazi et al., 2018).

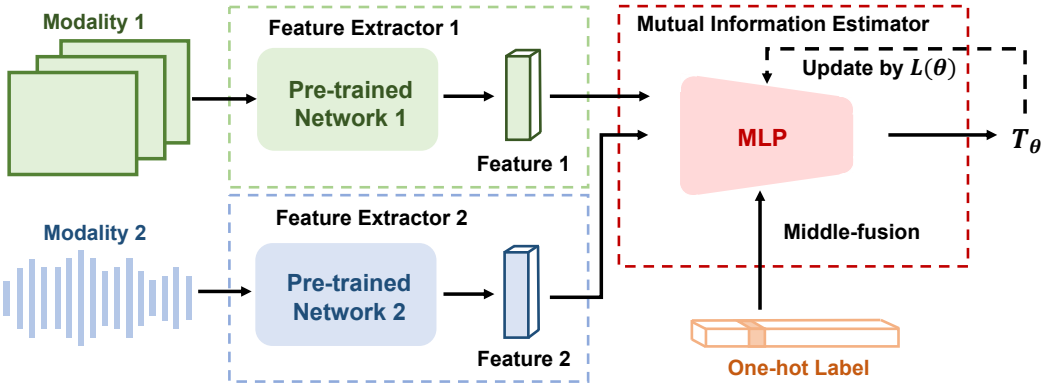

Figure 2: Pipeline to calculate the metric: First extract the features of the input data from two modalities by pre-trained models. Then apply the MINE to estimate the mutual information $I(Z;Y,X)$, $I(X;Y,Z)$, or $I(X;Z)$.

4.1 Metric Design

From the above analysis, it is natural to use $\Gamma_X + \Gamma_Z$ as the metric since they represent how much complementary information the modalities X and Z can provide exclusively about the target Y.

However, $\Gamma_X + \Gamma_Z$ is not enough for comparing among datasets. According to equation 1, the same amount of $\Gamma_X + \Gamma_Z$ could indicate different situations if the total information $I(X,Z;Y)$ is different. Therefore, to make the metric comparable among datasets, we need to perform normalization by dividing it with $I(X,Z;Y)$, written as

$$\frac{\Gamma_X + \Gamma_Z}{I(X,Z;Y)}$$

Now, our metric becomes the "proportion" of $\Gamma_X + \Gamma_Z$ in $I(X,Z;Y)$. When our metric is large, the modalities are more complementary to each other and more indispensable for the task. Note that this "proportion" could be greater than 1 because $I(X;Y;Z) = I(X,Z;Y) - \Gamma_X - \Gamma_Z$ may be negative. This happens when Z (or X) greatly increases the correlation strength between X (or Z) and Y. Without Z (or X), the other modality becomes nearly uncorrelated with the target Y. Hence, when the metric is greater than 1, it can still reflect the modality complementarity and reveals more about the interconnection between the modalities and the target.

4.2 Calculation

Now we consider how to calculate our metric. Γ_X and Γ_Z are in the form of conditional mutual information and could not be computed directly. We notice that

$$\Gamma_Z = I(Z;Y|X) = I(Z;Y,X) - I(Z;X)$$
$$\Gamma_X = I(X;Y|Z) = I(X;Y,Z) - I(X;Z)$$

So we transform the metric into

$$\frac{\Gamma_X + \Gamma_Z}{I(X,Z;Y)} = \frac{I(X;Y,Z) + I(Z;Y,X) - 2I(X;Z)}{I(X,Z;Y)}$$

Additionally, considering that most real-world datasets roughly satisfy the realizability assumption, i.e., there exists a function in the hypothesis space that can predict Y given X and Z with zero population risk, we could approximate $I(X,Z;Y) = H(Y) - H(Y|X,Z)$ with $H(Y)$ because the second term is close to zero. $H(Y)$ is easier to compute given the distribution of Y, especially when we focus on the classification task with discrete labels.

For each mutual information term with the form $I(A;B)$, we design a two-phase pipeline for computation (See Figure 2):

- In the first phase, we reduce the dimension of the high-dimensional input A and B to accelerate the computation by pre-trained feature extractors. The pre-trained models are shared among the calculation of all three terms.

- In the second phase, we use the extracted features as inputs for MINE (Belghazi et al., 2018) to compute the mutual information. Specifically, we calculate the value through optimization converging to a lower bound of the mutual information. For each iteration, we sample an m-sample batch $\{(\mathbf{a}^{(i)}, \mathbf{b}^{(i)})\}_{i=1}^{m}$ from the joint distribution $P(A, B)$ and an m-sample batch $\{\mathbf{b'}^{(i)}\}_{i=1}^{m}$ from the marginal distributions $P(B)$. Denote the estimator network as T and its parameters as θ. We evaluate the lower bound L as follows and the moving average of gradients of $L(\theta)$ for updating the network parameters.

$$L(\theta) = \frac{1}{m}\sum_{i=1}^{m} T_\theta(\mathbf{a}^{(i)}, \mathbf{b}^{(i)}) - \log(\frac{1}{m}\sum_{i=1}^{m}\exp(T_\theta(\mathbf{a}^{(i)}, \mathbf{b'}^{(i)})))$$

  We adjust the original MINE by adding the following trick: The calculation of $I(X; Y, Z)$ and $I(Z; Y, X)$ involve the target Y, so we concatenate the one-hot encoding label with the extracted feature in a middle-fusion fashion and ensure that the estimator network T could combine the two information sources. For more details, please see the supplementary material.

We believe that modality complementarity is crucial to the analysis of multi-modal robustness. Without controlling this factor, we cannot fairly compare experimental results on various multi-modal datasets, and thus we cannot derive a universal conclusion on multi-modal robustness. By calculating our metric on multi-modal datasets, we will better understand their difference in modality complementarity, leading to less biased comparisons and conclusions.

5 Experiments

We conduct experiments to verify the validity of our analysis and the effectiveness of our pipeline. We first introduce the training and testing settings and then show the results on the synthetic dataset, Additive AV-MNIST dataset, and real-world datasets. Unless otherwise specified, the missing/noise/adversarial robustness mentioned in the following subsections refers to the average accuracy under two sources of missing/noise/adversarial attack, divided by the model accuracy in the clean setting. For more detailed settings and results, please see the Appendix B and C.

Training setting. We use an MLP as the estimator. For different datasets, the structure of the MLP varies to match the input size. We train the estimator on the training set since in reality we only have access to it, and we assume that the validation set is i.i.d. sampled from the same distribution as the training set. For the pre-trained feature extractors, we use Resnet18 (He et al., 2015) and AudioNet (1-D CNN) (Tian & Xu, 2021) for Kinetics-Sounds and AVE, LeNet5 (LeCun et al., 1998) and an 2-D CNN for AAV-MNIST, and Resnet152 (He et al., 2015) and BERT (Devlin et al., 2019) for Hateful-Meme dataset. For the models tested for robustness, we use late fusion models for AVE, Kinetics-Sounds, and AV-MNIST, and we apply MMBT (Kiela et al., 2019) for Hateful-Meme dataset.

Test setting. We test the model robustness under two settings discussed above: missing modality and single noisy modality. We also explore the model robustness under adversarial attack. For the missing image or audio, we substitute them with the average of all inputs in the training set. For the missing text, we use a blank sentence <SOS><EOS> as the input. Note that the inputs are all scaled to the range $[-1, 1]$ (spectrogram) or $[0, 1]$ (image). For noisy image and audio, we add a Gaussian noise $N \sim \mathcal{N}(0, 0.5)$ to each dimension. For noisy text, we replace each word by a random word with a probability 0.5. For adversarial attack on image and audio, we use FGSM (Goodfellow et al., 2014) with step size $\epsilon = 0.03$. We use the results of missing text for adversarial text.

5.1 Synthetic Dataset

We first test our analysis on a well-designed synthetic dataset since we can adjust the degree of its modality complementarity. Inspired by previous work (Hessel & Lee, 2020) and (Huang et al., 2021), we generate a set of synthetic data (x, z, y). First, we sample random projection $P_X \in \mathbb{R}^{d_1 \times d}$, $P_Z \in \mathbb{R}^{d_2 \times d}$, and $P \in \mathbb{R}^{d \times d}$ from a uniform distribution $U(-0.5, 0.5)$. Then we repeat following steps:[1]

[1] We sample 5000 data points with 80/20 train/val split and $\langle d, d_1, d_2, \delta \rangle = \langle 50, 200, 100, 0.25 \rangle$.

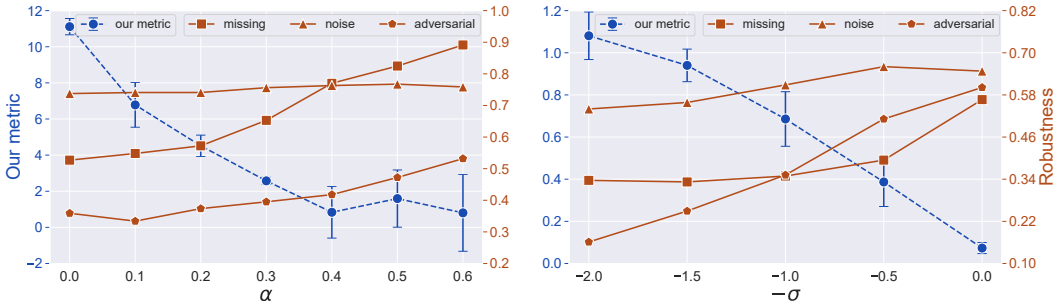

Figure 3: Two line plots showing the mean value of estimated metric (blue line) with error bars (standard variances) of three independent repeated experiments and tested robustness (orange line) on the synthetic dataset (left) and AAV-MNIST dataset (right). The x-axis is the parameter used in data generation. For the synthetic dataset, we plot α. For AAV-MNIST, we plot $-\sigma$ for unity. As the overlap of two modalities becomes larger, they are less complementary and our metric correspondingly goes down. Meanwhile, the tested robustness increases in all three settings. The variance and trend of each mutual information term can be found in the supplementary material.

Step 1. Sample $x, z \in \mathbb{R}^d \sim \mathcal{N}(0, 1)$.

Step 2. Set $z \leftarrow (1 - \alpha)z + \alpha x$; Then do projection $z \leftarrow Pz$.

Step 3. Normalize x, z to unit length; if $|x \cdot z| \leq \delta$, return to the Step 2.

Step 4. Generate the label y: If $x \cdot z > 0$, then $y = 1$; else $y = 0$. Return the tuple $(P_X x, P_Z z, y)$.

The α used in data generation controls the overlap between the two modalities X, Z. When $\alpha = 0$, the modalities are independent and complementary in predicting the label Y. When $\alpha = 1$, they are redundant for prediction. Viewing the synthetic dataset with different α as different datasets, we calculate our metric using the pipeline in section 4 and test the robustness of simple two-layer perceptron neural networks [2] trained on these datasets. The results are shown in the plot 3. In each dataset, our pipeline can estimate the proposed metric and quantify the complementarity of the two modalities. Further, the model robustness decreases as the complementarity increases, which verifies our analysis.

5.2 Additive AV-MNIST

To show that our pipeline can effectively estimate the modality complementarity of more complex datasets, we further design a toy dataset named Additive AV-MNIST (AAV-MNIST) adapted from the AV-MNIST dataset (Vielzeuf et al., 2018). The modality complementarity can be controlled by a parameter σ in the data generation process. Below, we show how to generate AAV-MNIST dataset from the original AV-MNIST dataset. The following steps are repeated for every image i in AV-MNIST:

Step 1. Let x be the label of i. Sample $\delta \in \mathbb{R} \sim \mathcal{N}(0, \sigma)$ and round δ to the nearest integer.

Step 2. Set $y \leftarrow (x + \delta) \mod 10$. Uniformly sample a spectrogram s from all spectrograms in AV-MNIST with label y.

Step 3. Calculate the new label $t \leftarrow (x + y)/2$. Round t to the nearest integer. Return the tuple (i, s, t).

The AAV-MNIST dataset is an extension of AV-MNIST dataset. When $\sigma = 0$, AAV-MNIST dataset is equivalent to AV-MNIST dataset where each image and its paired spectrogram represent the same number. As σ increases, each image become less correlated with its paired spectrogram, so they become more complementary for predicting the label t.

We show in the plot 3 that our metric reflects the complementarity of the AAV-MNIST dataset with different σ, indicating that our pipeline is effective in more complex settings beyond the synthetic

[2] Each achieves accuracy $> 96\%$ on corresponding validation sets. The neural network structure and more training settings can be found in the supplementary materials.

Dataset	Our metric	Clean	Missing	Noisy	Adversarial
AAV-MNIST($\sigma = 2.0$)	0.9212	0.6435	0.3368	0.5399	0.1612
Hateful-Meme	0.2403	0.3249	0.1005	0.5171	0.3144
AV-MNIST	0.0490	0.9969	0.5666	0.6478	0.6012
Kinetics-Sounds	0.0455	0.6387	0.5540	0.6098	0.2672
AVE	0.0126	0.7637	0.4838	0.5831	0.3355

Table 1: Our estimated metric and tested robustness of real-world datasets: Kinetics-Sounds, AVE, AV-MNIST, and Hateful-Meme. Since the Hateful-Meme Challenge is a binary classification task, we use F1 score for evaluation instead of accuracy. We also provide results in clean setting for reference.

dataset. Further, the robustness in the three settings verifies our conclusion that with other conditions unchanged, the more complementary the modalities are, the less robust the best model will be.

5.3 Real-world Datasets

Now we apply our pipeline to real-world datasets to investigate their modality complementarity. Our results on AVE (Tian et al., 2018), Kinetics-Sounds (Carreira & Zisserman, 2017; Arandjelovic & Zisserman, 2017), and Hateful-Meme dataset (Kiela et al., 2020a) are shown in the table 1. The details of these datasets are described in the Appendix B. We also list results on the AV-MNIST dataset and AAV-MNIST ($\sigma = 2.0$) for reference.

The low value in our metric of AVE, Kinetics-Sounds, and AV-MNIST indicates that they possess relatively little modality complementarity, revealing the heavy redundancy between the two modalities. On the contrary, the modalities in the Hateful-Meme dataset are more complementary. This finding suits our intuition: In the Hateful-Meme dataset, altering the paired text of an image probably changes the label (Kiela et al., 2020b). Hence, only perceiving the image would not derive the right answer. For the event classification task defined by Kinetics-Sounds or AVE, the audio and frames both lead to a rough answer.

The tested robustness demonstrates how the modality complementarity affects model robustness. The missing case affects AAV-MNIST($\sigma = 2.0$) and Hateful-Meme far more than the other three datasets. They are also more vulnerable in single source noisy case than other datasets. Hence, to compare model robustness among these datasets, we should take modality complementarity into account. For instance, we only compare robustness among datasets with a similar degree of modality complementarity, or we can normalize the results by our metric. We show an analysis of the existing measure of modality missing robustness by applying our metric in the Appendix C.3.

Furthermore, the model robustness, especially the adversarial robustness, is also affected by factors other than modality complementarity. For instance, the model adversarial robustness of AVE and Kinetics-Sounds dataset is significantly lower than that of AV-MNIST dataset. We conjecture that this is related to the number of robust features in each modality of the datasets, which requires future work to confirm.

6 Conclusions

In this work, we partly explain the contradiction in previous conclusions on multi-modal robustness by pointing out the importance of the modality complementarity through information-theoretical analysis and carefully-designed experiments. As a reflection of modality interconnection, our proposed metric provides a basis for better understanding various multi-modal datasets/tasks and can be used beyond analyzing multi-modal robustness.

Reproducibility Statement

We provide the source code and configuration for the key experiments, including instructions on generating data, training the models, and evaluating the robustness. We thoroughly checked the code implementations and empirically verified the effectiveness of our method. All proofs are stated in the appendix with explanations and underlying assumptions.

REFERENCES

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4971–4980, 2018.

Massih R. Amini, Nicolas Usunier, and Cyril Goutte. Learning from multiple partially observed views - an application to multilingual text categorization. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta (eds.), *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009. URL https://proceedings.neurips.cc/paper/2009/file/f79921bbae40a577928b76d2fc3edc2a-Paper.pdf.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. *CoRR*, abs/1505.00468, 2015. URL http://arxiv.org/abs/1505.00468.

Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. *CoRR*, abs/1705.08168, 2017. URL http://arxiv.org/abs/1705.08168.

Philip Bachman, R. Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *CoRR*, abs/1906.00910, 2019. URL http://arxiv.org/abs/1906.00910.

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *CoRR*, abs/1705.09406, 2017. URL http://arxiv.org/abs/1705.09406.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 531–540. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/belghazi18a.html.

Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian J. Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *CoRR*, abs/1902.06705, 2019. URL http://arxiv.org/abs/1902.06705.

João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. *CoRR*, abs/1705.07750, 2017. URL http://arxiv.org/abs/1705.07750.

Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. *Advances in neural information processing systems*, 28, 2015.

Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. *CoRR*, abs/1902.02918, 2019. URL http://arxiv.org/abs/1902.02918.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *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)*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1423. URL https://doi.org/10.18653/v1/n19-1423.

Daisy Yi Ding, Balasubramanian Narasimhan, and Robert Tibshirani. Cooperative learning for multi-view analysis, 2021. URL https://arxiv.org/abs/2112.12337.

Andreas Eitel, Jost Tobias Springenberg, Luciano Spinello, Martin A. Riedmiller, and Wolfram Burgard. Multimodal deep learning for robust RGB-D object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2015, Hamburg, Germany, September 28 - October 2, 2015*, pp. 681–687. IEEE, 2015. doi: 10.1109/IROS.2015.7353446. URL https://doi.org/10.1109/IROS.2015.7353446.

M. Feder and N. Merhav. Relations between entropy and error probability. *IEEE Transactions on Information Theory*, 40(1):259–266, 1994. doi: 10.1109/18.272494.

Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=B1xwcyHFDr.

Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. *CoRR*, abs/1604.06573, 2016a. URL http://arxiv.org/abs/1604.06573.

Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1933–1941, 2016b.

Nic Ford, Justin Gilmer, Nicholas Carlini, and Ekin Dogus Cubuk. Adversarial examples are a natural consequence of test error in noise. *CoRR*, abs/1901.10513, 2019. URL http://arxiv.org/abs/1901.10513.

Keinosuke Fukunaga and Donald M Hummels. Bayes error estimation using parzen and k-nn procedures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (5):634–643, 1987.

Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. *CoRR*, abs/1912.04487, 2019. URL http://arxiv.org/abs/1912.04487.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2014. URL https://arxiv.org/abs/1412.6572.

Jiuxiang Gu, Jianfei Cai, Shafiq R. Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. *CoRR*, abs/1711.06420, 2017. URL http://arxiv.org/abs/1711.06420.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. *CoRR*, abs/2102.02051, 2021. URL https://arxiv.org/abs/2102.02051.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *CoRR*, abs/1903.12261, 2019. URL http://arxiv.org/abs/1903.12261.

Jack Hessel and Lillian Lee. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! *CoRR*, abs/2010.06572, 2020. URL https://arxiv.org/abs/2010.06572.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization, 2018. URL https://arxiv.org/abs/1808.06670.

Ruitong Huang, Bing Xu, Dale Schuurmans, and Csaba Szepesvári. Learning with a strong adversary. *CoRR*, abs/1511.03034, 2015. URL http://arxiv.org/abs/1511.03034.

Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multimodal learning better than single (provably). *CoRR*, abs/2106.04538, 2021. URL https://arxiv.org/abs/2106.04538.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). *arXiv preprint arXiv:2203.12221*, 2022.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. *CoRR*, abs/1612.06890, 2016. URL http://arxiv.org/abs/1612.06890.

Sham M Kakade and Dean P Foster. Multi-view regression via canonical correlation analysis. In *International Conference on Computational Learning Theory*, pp. 82–96. Springer, 2007.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text. *CoRR*, abs/1909.02950, 2019. URL http://arxiv.org/abs/1909.02950.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *CoRR*, abs/2005.04790, 2020a. URL https://arxiv.org/abs/2005.04790.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *arXiv preprint arXiv:2005.04790*, 2020b.

Taewan Kim and Joydeep Ghosh. On single source robustness in deep fusion models. *CoRR*, abs/1906.04691, 2019a. URL http://arxiv.org/abs/1906.04691.

Taewan Kim and Joydeep Ghosh. On single source robustness in deep fusion models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019b. URL https://proceedings.neurips.cc/paper/2019/file/8420d359404024567b5aefda1231af24-Paper.pdf.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Linjie Li, Zhe Gan, and Jingjing Liu. A closer look at the robustness of vision-and-language pre-trained models. *CoRR*, abs/2012.08673, 2020. URL https://arxiv.org/abs/2012.08673.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Yufan Chen, Peter Wu, Michelle A Lee, Yuke Zhu, et al. Multibench: Multiscale benchmarks for multimodal representation learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions, 2022. URL https://arxiv.org/abs/2209.03430.

Yijie Lin, Yuanbiao Gou, Zitao Liu, Boyun Li, Jiancheng Lv, and Xi Peng. COMPLETER: incomplete multi-view clustering via contrastive prediction. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 11174–11183. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.01102. URL https://openaccess.thecvf.com/content/CVPR2021/html/Lin_COMPLETER_Incomplete_Multi-View_Clustering_via_Contrastive_Prediction_CVPR_2021_paper.html.

Yijie Lin, Yuanbiao Gou, Xiaotian Liu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. Dual contrastive prediction for incomplete multi-view representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–14, 2022. doi: 10.1109/TPAMI.2022.3197238.

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

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality?, 2022.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2017. URL https://arxiv.org/abs/1706.06083.

Kaleel Mahmood, Rigel Mahmood, and Marten van Dijk. On the robustness of vision transformers to adversarial examples. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 7818–7827. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00774. URL https://doi.org/10.1109/ICCV48922.2021.00774.

David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. *CoRR*, abs/1811.04251, 2018. URL http://arxiv.org/abs/1811.04251.

Dongyu Meng and Hao Chen. Magnet: a two-pronged defense against adversarial examples. *CoRR*, abs/1705.09064, 2017. URL http://arxiv.org/abs/1705.09064.

Nicolas Papernot, Patrick D. McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. *CoRR*, abs/1511.04508, 2015. URL http://arxiv.org/abs/1511.04508.

Kun Qian, Shilin Zhu, Xinyu Zhang, and Li Erran Li. Robust multimodal vehicle detection in foggy weather using complementary lidar and radar signals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 444–453, June 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. URL https://arxiv.org/abs/2103.00020.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *CoRR*, abs/2102.12092, 2021. URL https://arxiv.org/abs/2102.12092.

Marco Ramoni and Paola Sebastiani. Robust learning with missing data. *Machine Learning*, 45: 147–170, 11 2001. doi: 10.1023/A:1010968702992.

Daniel Rosenberg, Itai Gat, Amir Feder, and Roi Reichart. Are VQA systems rad? measuring robustness to augmented data with focused interventions. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pp. 61–70. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-short.10. URL https://doi.org/10.18653/v1/2021.acl-short.10.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *CoRR*, abs/1703.00810, 2017. URL http://arxiv.org/abs/1703.00810.

Carl-Johann Simon-Gabriel, Yann Ollivier, Leon Bottou, Bernhard Schölkopf, and David Lopez-Paz. First-order adversarial vulnerability of neural networks and input dimension. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5809–5817. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/simon-gabriel19a.html.

Vikas Sindhwani, Partha Niyogi, and Mikhail Belkin. A co-regularization approach to semi-supervised learning with multiple views. In *Proceedings of ICML workshop on learning with multiple views*, volume 2005, pp. 74–79. Citeseer, 2005.

Amanpreet Singh, Vedanuj Goswami, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Mmf: A multimodal framework for vision and language research. https://github.com/facebookresearch/mmf, 2020.

Karthik Sridharan and Sham M. Kakade. An information theoretic framework for multi-view learning. In Rocco A. Servedio and Tong Zhang (eds.), *21st Annual Conference on Learning Theory - COLT 2008, Helsinki, Finland, July 9-12, 2008*, pp. 403–414. Omnipress, 2008. URL http://colt2008.cs.helsinki.fi/papers/94-Sridharan.pdf.

William H Sumby and Irwin Pollack. Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america*, 26(2):212–215, 1954.

Xinwei Sun, Yilun Xu, Peng Cao, Yuqing Kong, Lingjing Hu, Shanghang Zhang, and Yizhou Wang. TCGM: an information-theoretic framework for semi-supervised multi-modality learning. *CoRR*, abs/2007.06793, 2020. URL https://arxiv.org/abs/2007.06793.

Yapeng Tian and Chenliang Xu. Can audio-visual integration strengthen robustness under multimodal attacks? *CoRR*, abs/2104.02000, 2021. URL https://arxiv.org/abs/2104.02000.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. *CoRR*, abs/1803.08842, 2018. URL http://arxiv.org/abs/1803.08842.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *CoRR*, abs/1906.05849, 2019. URL http://arxiv.org/abs/1906.05849.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method, 2000. URL https://arxiv.org/abs/physics/0004057.

Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive learning, multi-view redundancy, and linear models. In Vitaly Feldman, Katrina Ligett, and Sivan Sabato (eds.), *Algorithmic Learning Theory, 16-19 March 2021, Virtual Conference, Worldwide*, volume 132 of *Proceedings of Machine Learning Research*, pp. 1179–1206. PMLR, 2021. URL http://proceedings.mlr.press/v132/tosh21a.html.

Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. Missing modalities imputation via cascaded residual autoencoder. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4971–4980, 2017. doi: 10.1109/CVPR.2017.528.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. *CoRR*, abs/1806.06176, 2018. URL http://arxiv.org/abs/1806.06176.

Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Demystifying self-supervised learning: An information-theoretical framework. *CoRR*, abs/2006.05576, 2020. URL https://arxiv.org/abs/2006.05576.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. URL http://arxiv.org/abs/1807.03748.

Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. Centralnet: a multilayer approach for multimodal fusion. *CoRR*, abs/1808.07275, 2018. URL http://arxiv.org/abs/1808.07275.

Jinghua Wang, Zhenhua Wang, Dacheng Tao, Simon See, and Gang Wang. Learning common and specific features for rgb-d semantic segmentation with deconvolutional networks. In *European Conference on Computer Vision*, pp. 664–679. Springer, 2016.

Shaojie Wang, Tong Wu, and Yevgeniy Vorobeychik. Towards robust sensor fusion in visual perception. *CoRR*, abs/2006.13192, 2020. URL https://arxiv.org/abs/2006.13192.

Yihong Wu and Sergio Verdu. Functional properties of minimum mean-square error and mutual information. *IEEE Transactions on Information Theory*, 58(3):1289–1301, 2012. doi: 10.1109/TIT.2011.2174959.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

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

Karren Yang, Wan-Yi Lin, Manash Barman, Filipe Condessa, and Zico Kolter. Defending multimodal fusion models against single-source adversaries. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3339–3348, 2021. doi: 10.1109/CVPR46437.2021.00335.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: collision events for video representation and reasoning. *CoRR*, abs/1910.01442, 2019. URL http://arxiv.org/abs/1910.01442.

Youngjoon Yu, Hong Joo Lee, Byeong Cheon Kim, Jung Uk Kim, and Yong Man Ro. Investigating vulnerability to adversarial examples on multimodal data fusion in deep learning. *CoRR*, abs/2005.10987, 2020. URL https://arxiv.org/abs/2005.10987.

Lei Yuan, Yalin Wang, Paul M. Thompson, Vaibhav A. Narayan, and Jieping Ye. Multi-source learning for joint analysis of incomplete multi-modality neuroimaging data. In Qiang Yang, Deepak Agarwal, and Jian Pei (eds.), *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012*, pp. 1149–1157. ACM, 2012. doi: 10.1145/2339530.2339710. URL https://doi.org/10.1145/2339530.2339710.

Changqing Zhang, Zongbo Han, yajie cui, Huazhu Fu, Joey Tianyi Zhou, and Qinghua Hu. Cpm-nets: Cross partial multi-view networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019a. URL https://proceedings.neurips.cc/paper/2019/file/11b9842e0a271ff252c1903e7132cd68-Paper.pdf.

Wenwei Zhang, Hui Zhou, Shuyang Sun, Zhe Wang, Jianping Shi, and Chen Change Loy. Robust multi-modality multi-object tracking. *CoRR*, abs/1909.03850, 2019b. URL http://arxiv.org/abs/1909.03850.

Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 570–586, 2018.

Stephan Zheng, Yang Song, Thomas Leung, and Ian J. Goodfellow. Improving the robustness of deep neural networks via stability training. *CoRR*, abs/1604.04326, 2016. URL http://arxiv.org/abs/1604.04326.

A PROOFS OF RESULTS IN SECTION 3

Below, we articulate the proof omitted in section 3.

A.1 PROOF OF THEOREM 3.1

Proof. We leverage the previous results in (Feder & Merhav, 1994) and (Cover, 1999):

$$-\log(1-P_{e_c}) \leq H(Y|X,Z)$$
$$H(Y|X,Z) \leq \log 2 + P_{e_c}\log|Y|.$$

Combine the two inequalities and put P_{e_c} in the middle:

$$\frac{H(Y|X,Z)-\log 2}{\log|Y|} \leq P_{e_c} \leq 1-\exp(-H(Y|X,Z))$$

which is the first result in the theorem. Then we apply the results to $P_{e_c}^{Miss}$:

$$\frac{H(Y|X)-\log 2}{\log|Y|} \leq P_{e_c}^{Miss} \leq 1-\exp(-H(Y|X)).$$

Since

$$\begin{aligned}\Gamma_Z &= I(Z;Y|X)\\ &= I(Y;X,Z)-I(Y;X)\\ &= [H(Y)-H(Y|X,Z)]-[H(Y)-H(Y|X)]\\ &= H(Y|X)-H(Y|X,Z)\end{aligned}$$

plug Γ_Z into the result above, and we derive

$$\frac{H(Y|X,Z)+\Gamma_Z-\log 2}{\log|Y|} \leq P_{e_c}^{Miss} \leq 1-\exp(-H(Y|X,Z)-\Gamma_Z)$$

which is the second result in the theorem.

Now we consider the case when Y is continuous. Based on the orthogonality principle and Theorem 10 in (Wu & Verdu, 2012), if $Y \in [-1,1]$, we have

$$\begin{aligned}P_{e_r}^{\text{Miss}} - P_{e_r} &= \mathbb{E}_{x,y\sim P_{X,Y}}[(y-\mathbb{E}[Y|x])^2] - \mathbb{E}_{x,z,y\sim P_{X,Z,Y}}[(y-\mathbb{E}[Y|x,z])^2]\\ &= \mathbb{E}_{x,z,y\sim P_{X,Z,Y}}[(\mathbb{E}[Y|x]-\mathbb{E}[Y|x,z])^2]\\ &\leq \frac{1}{2}I(Z;Y|X)\end{aligned}$$

which is the third result in the theorem. □

A.2 PROOF OF PROPOSITION 1

Proof.

$$\begin{aligned}I(Z;R_N(Z)) &= H(R_N(Z)) - H(R_N(Z)|Z)\\ &= H(Z+N) - H(Z+N|Z)\\ &= H(Z+N) - H(N|Z)\\ &= H(Z+N) - H(N)\\ &= H(Z+N) - \frac{1}{2}\log(2\pi e\sigma)\end{aligned}$$

since $N \sim \mathcal{N}(0,\sigma)$. Now we consider $H(Z+N)$. Because

$$E[(Z+N)^2] = E[Z^2] + 2E[ZN] + E[N^2] \leq p_Z + \sigma.$$

Therefore

$$H(Z+N) \leq \frac{1}{2}\log(2\pi e(\sigma+p_Z))$$

and thus

$$I(Z;R_N(Z)) = H(Z+N) - \frac{1}{2}\log(2\pi e\sigma) \leq \frac{1}{2}\log(\frac{\sigma+p_Z}{\sigma}) = \frac{1}{2}\log(1+\frac{p_Z}{\sigma}).$$

□

A.3 PROOF OF THEOREM 3.2

Proof. From the first inequality 2 in Theorem 3.1, we obtain

$$P_{e_c} \geq \frac{H(Y|X,Z) - \log 2}{\log |Y|}.$$

In the noisy case, we have

$$P_{e_c}^{No} \geq \frac{H(Y|X,R_N(Z)) - \log 2}{\log |Y|}.$$

Note that $R_N(Z) \perp Y|Z$. So $H(R_N(Z)|Z) = H(R_N(Z)|Z,Y) \leq H(R_N(Z)|Y)$ and

$$I(Y;R_N(Z)) = H(R_N(Z)) - H(R_N(Z)|Y) \leq H(R_N(Z)) - H(R_N(Z)|Z) = I(Z;R_N(Z)).$$

The difference between $H(Y|X,R_N(Z))$ and $H(Y|X,Z)$ has following lower bound:

$$\begin{aligned}
H(Y|X,R_N(Z)) - H(Y|X,Z) &= H(Y|X,R_N(Z)) - H(Y|X,Z,R_N(Z)) \\
&= I(Y;Z|X,R_N(Z)) \\
&= I(Y;Z,R_N(Z)|X) - I(Y;R_N(Z)|X) \\
&= I(Y;Z|X) - I(Y;R_N(Z)|X) \\
&= I(Y;Z|X) - I(Y;R_N(Z)) + I(Y;R_N(Z);X) \\
&\geq I(Y;Z|X) - I(Z;R_N(Z)) + I(Y;R_N(Z);X) \\
&\geq \Gamma_Z - \frac{1}{2}\log(1+\frac{p_Z}{\sigma}) + I(Y;R_N(Z);X).
\end{aligned}$$

Plug this result in and we obtain

$$\begin{aligned}
P_{e_c}^{No} &\geq \frac{H(Y|X,Z) + \Gamma_Z - \frac{1}{2}\log(1+\frac{p_Z}{\sigma}) + I(Y;R_N(Z);X) - \log 2}{\log |Y|} \\
&= \frac{H(Y|X,Z) + \Gamma_Z + I(X;Y;R_N(Z)) - \frac{1}{2}\log(4+\frac{4p_Z}{\sigma})}{\log |Y|}.
\end{aligned}$$

□

B EXPERIMENT DETAILS

In this section, we introduce the details of experiment settings. For each estimator, we adopt the bias correction method in MINE (Belghazi et al., 2018). When calculating gradients, we replace the denominator of the second term (the log-term) by its exponential moving average with an update rate $= 0.01$. To lower the variance of our estimation, we report the average result of the last 100 batches.

B.1 SYNTHETIC DATASET

The data generation method of synthetic data is described in section 5. Note that each class contains exact half of all the data points. For synthetic dataset, we do not employ feature extractors. We directly use the 200-dim X and 100-dim Z as the input for the estimators which are designed as following MLPs (See table 2). We use enough parameters to ensure the expressivity of the estimator. We train the estimators for 500 epochs and apply batch size $= 100$. We use Adam as the optimizer, 10^{-3} as the learning rate, and $2 \cdot 10^{-4}$ as the weight decay rate.

For robustness testing, we trained a simple two-layer perceptron neural network. In the missing modality setting, we fill in the missing modality with zeros. In the noisy setting, we add Gaussian noise $\mathcal{N}(0,1)$ to the specified modality. In the adversarial setting, we update the inputs of the specific modality according to the gradients' sign where we can maximize the loss. And we set 0.1 as the step size.

Layer	# of Parameters
linear1	$300 \times 1000 + 1000$
linear2	$1000 \times 500 + 500$
linear3	$500 \times 100 + 100$
linear4	$100 \times 1 + 1$

Layer	# of Parameters
linear1	$300 \times 1000 + 1000$
linear2	$1000 \times 200 + 200$
linear3	$200 \times 10 + 10$
linear4	$12 \times 12 + 12$
linear5	$12 \times 1 + 1$

Layer	# of Parameters
linear1	$300 \times 200 + 200$
linear2	$200 \times 2 + 2$

Table 2: At left are the numbers of parameters in each layer for the estimator of $I(X;Z)$ for synthetic dataset. The activation function for linear1$\sim$3 is ELU and linear4 has no activation function. In the middle are the numbers of parameters in each layer for the estimator of $I(X;Z,Y)$ and $I(Z;X,Y)$ for synthetic dataset. The activation function for linear1$\sim$4 is ELU while linear5 has no activation function. The one-hot label is concatenated with the output of linear3 and then input to linear4. At right are the numbers of parameters in each layer of the trained two-layer perceptron for testing robustness.

B.2 AAV-MNIST Dataset

The AV-MNIST dataset (Vielzeuf et al., 2018) contains two modalities: Disturbed images and audio spectrograms. The image is from the 28×28 PCA-projected MNIST images with 75% energy removed. The 112×112 audio spectrograms are the pronounced digits augmented with noise samples. The size of the AV-MNIST dataset is the same as that of the MNIST dataset (with train/test splits of 55000/10000).

For AAV-MNIST dataset, the data generation method is also in section 5. So the size of the AAV-MNIST dataset is the same as that of AV-MNIST dataset. To magnify the contrast between AAV-MNIST with different σ, we alter the energy eliminating rate in the original AV-MNIST dataset by reducing it from 75% to 25%. In this way, the image can be more indicative of the target Y. We do not use data augmentation for AAV-MNIST, except that the inputs are all scaled to the range $[-1, 1]$ (spectrogram) or $[0, 1]$ (image). We employ pre-trained models for AV-MNIST as feature extractors. Specifically, we first train a LeNet5 (LeCun et al., 1998) for visual classification on AV-MNIST and also a 2D-CNN for audio classification on AV-MNIST. Then we use the output of the penultimate layer (before the final linear layer) as features, 84-dim for each modality. The estimator is designed as follows (See table 3).

Layer	# of Parameters
linear1	$168 \times 1000 + 1000$
linear2	$1000 \times 100 + 100$
linear3	$100 \times 1 + 1$

Layer	# of Parameters
linear1	$168 \times 1000 + 1000$
linear2	$1000 \times 100 + 100$
linear3	$110 \times 110 + 110$
linear4	$110 \times 1 + 1$

Table 3: At left are the numbers of parameters in each layer for the estimator of $I(X;Z)$ for AAV-MNIST dataset. The activation function for linear1$\sim$2 is ELU and linear3 has no activation function. At right are the numbers of parameters in each layer for the estimator of $I(X;Z,Y)$ and $I(Z;X,Y)$ for AAV-MNIST dataset. The activation function for linear1$\sim$3 is ELU while linear4 has no activation function. The one-hot label is concatenated with the output of linear2 and then input to linear3.

We train the estimators for 50 epochs and apply batch size $= 800$. We set Adam as the optimizer, 10^{-3} as the learning rate for the first forty epochs and 10^{-4} for the last ten epochs, and 10^{-4} as the weight decay rate.

We test the robustness of a late-fusion model for classification. It uses the pre-trained LeNet5 and 2D-CNN as feature extractors, concatenates their outputs, and uses a linear layer as the fusion layer. During training, the feature extractors and the fusion layer are trained jointly.

In the missing modality setting, we fill in the missing modality with the average of data of this modality in the training set. As for noisy settings, we add Gaussian noise ($\mathcal{N}(0, 0.5)$) to the specified modality. When testing adversarial robustness, we adopt the FGSM implementation from (Tian & Xu, 2021). We do not test robustness under other attacks because more effective attacks can reduce the accuracy to near zero and thus make it hard for comparison.

B.3 KINETICS-SOUNDS AND AVE DATASET

The AVE dataset (Tian et al., 2018) contains 4,143 unconstrained videos spanning 28 event categories including Rodents, Accordion, and Mandolin. It has the train/val/test splits of 3,339/402/402 videos. The Kinetics-Sounds dataset (Arandjelovic & Zisserman, 2017; Carreira & Zisserman, 2017) consists of YouTube videos with manually annotated human actions. It is a subset with 27 classes from Kinetics400, such as playing_harmonica, tapping_pen, and shoveling_snow. It has the train/val/test splits of 9309/3104/3103.

For each dataset, we adopt the data processing and augmentation used in (Tian & Xu, 2021). Note that the inputs are all scaled to the range $[-1, 1]$ (spectrogram) or $[0, 1]$ (image). We train a Resnet18 (He et al., 2015) for visual classification and AudioNet (1-D CNN) (Tian & Xu, 2021) for audio classification. Then we use the output of the penultimate layer (before the final linear layer) as features, 512-dim for each modality. The estimator is designed as follows (See table 4).

Layer	# of Parameters
linear1	$1024 \times 1000 + 1000$
linear2	$1000 \times 100 + 100$
linear3	$100 \times 1 + 1$

Layer	# of Parameters
linear1	$1024 \times 1000 + 1000$
linear2	$1000 \times 100 + 100$
linear3	$128 \times 128 + 128$
linear4	$128 \times 1 + 1$

Table 4: At left are the numbers of parameters in each layer for the estimator of $I(X; Z)$ for AVE and Kinetics-Sounds dataset. The activation function for linear1$\sim$2 is ELU and linear3 has no activation function. At right are the numbers of parameters in each layer for the estimator of $I(X; Z, Y)$ and $I(Z; X, Y)$ for AVE dataset. For Kinetics-Sounds, linear3 has $127 \times 127 + 127$ parameters and linear4 has $127 \times 1 + 1$ parameters. The activation function for linear1$\sim$3 is ELU while linear4 has no activation function. The one-hot label is concatenated with the output of linear2 and then input to linear3.

For AVE dataset, we train the estimators for 30 epochs and apply batch size $= 96$. We use Adam as optimizer, 10^{-3} as learning rate and 10^{-4} as weight decay rate. For Kinetics-Sounds dataset, we train the estimators for 30 epochs and apply batch size $= 384$. We use Adam as the optimizer, $10^{-2}, 10^{-3}, 10^{-4}$ as the learning rate for 0$\sim$10, 10$\sim$20, and 20$\sim$30 epochs, respectively. We set 10^{-5} as the weight decay rate.

In practice, when estimating mutual information between A and B, we found that when the correlation between A and B is small, a larger learning rate should be used. However, a too-large learning rate quickly leads to an explosion due to the exponential term. The batch size should also be relatively small to increase the update frequency.

We test the robustness of a late-fusion model with a linear fusion layer. During training, the feature extractors and the fusion layer are trained jointly. When testing adversarial robustness, we adopt the FGSM implementation from (Tian & Xu, 2021). Settings for other robustness testing are in section 5.

B.4 HATEFUL-MEME DATASET

The Hateful-Meme dataset defines the task of identifying hate speech in memes. It has two classes of image-text pairs: hateful and benign. The same image with different captions can have different labels, so this dataset is expected to have a high degree of modal complementarity. It has the train/val/test splits of 8500/500/1000.

We adopt the data processing and augmentation used in (Singh et al., 2020). We use a Resnet152 (He et al., 2015) for visual input and BERT (Devlin et al., 2019) for texts. Then we extract the 2048-dimensional visual features and 768-dimensional language features, and use an estimator for calculating mutual information. The estimator is designed as follows (See table 5).

For Hateful-Meme dataset, we train the estimators for 30 epochs and apply batch size $= 128$. We use AdamW as the optimizer, 10^{-2} as the learning rate and 10^{-8} as the epsilon.

Layer	# of Parameters
linear1	$2816 \times 1000 + 1000$
linear2	$1000 \times 100 + 100$
linear3	$100 \times 1 + 1$

Layer	# of Parameters
linear1	$2816 \times 1000 + 1000$
linear2	$1000 \times 100 + 100$
linear3	$102 \times 102 + 102$
linear4	$102 \times 1 + 1$

Table 5: At left are the numbers of parameters in each layer for the estimator of $I(X;Z)$ for Hateful-Meme dataset. The activation function for linear1$\sim$2 is ELU and linear3 has no activation function. At right are the numbers of parameters in each layer for the estimator of $I(X;Z,Y)$ and $I(Z;X,Y)$ for Hateful-Meme dataset. The activation function for linear1$\sim$3 is ELU while linear4 has no activation function. The one-hot label is concatenated with the output of linear2 and then input to linear3.

We test the robustness of a MMBT model (Kiela et al., 2019). It uses the pre-trained Resnet152 and BERT as feature extractors. During training, the feature extractors and the fusion layer are trained jointly. Settings for robustness testing are described in section 5.

C More Experiment Results

In this section, we provide the numerical results and the line plots and analysis for each mutual information term.

C.1 Numerical Results

The numerical results for synthetic dataset and AAV-MNIST dataset are listed in the table 6. Since the robustness is affected by other factors, the robustness shown in the table is not monotonically increasing along with the decline of our metric. When we control other factors by only comparing datasets of the same kind but with a different parameter for generation, we observe that the robustness increases.

Dataset	Our metric	Missing	Noisy	Adversarial
Synthetic($\alpha = 0.0$)	10.738	0.5269	0.7375	0.3589
Synthetic($\alpha = 0.1$)	7.8197	0.5480	0.7409	0.3340
Synthetic($\alpha = 0.2$)	4.9031	0.5722	0.7409	0.3735
Synthetic($\alpha = 0.3$)	2.8570	0.6528	0.7561	0.3951
Synthetic($\alpha = 0.4$)	0.5598	0.7701	0.7629	0.4178
Synthetic($\alpha = 0.5$)	1.3624	0.8247	0.7674	0.4721
Synthetic($\alpha = 0.6$)	0.3702	0.8913	0.7585	0.5316
AAV-MNIST($\sigma = 2.0$)	0.9212	0.3367	0.5399	0.1612
AAV-MNIST($\sigma = 1.5$)	0.8307	0.3322	0.5584	0.2493
AAV-MNIST($\sigma = 1.0$)	0.5028	0.3489	0.6084	0.3520
AAV-MNIST($\sigma = 0.5$)	0.2224	0.3944	0.6605	0.5114
AAV-MNIST($\sigma = 0.0$)	0.0489	0.5665	0.6477	0.6012

Table 6: Our tested robustness and estimated metric on synthetic datasets and AAV-MNIST dataset. Recall that missing/noise/adversarial robustness refers to the average accuracy under two missing/noise/adversarial attack sources, divided by the model accuracy in the clean setting.

In addition, we provide the accuracy on following datasets in the clean setting for reference. See table 7.

C.2 The Change of Mutual Information

The change of each mutual information term in the calculation of our metric is shown in the plots in Figure 4. As the overlap of two modalities becomes larger, the amount of three mutual information terms grows with different rates. $I(X;Z)$ grows from nearly zero to large values, while $I(X;Y,Z)$ and $I(Z;Y,X)$ starts from non-zero values to larger values. The increasing amount of $I(X;Z)$ is

Dataset	uni-modal-X	uni-modal-Z	multi-modal
Synthetic($\alpha = 0.0$)	0.502	0.535	0.964
Synthetic($\alpha = 0.1$)	0.535	0.541	0.967
Synthetic($\alpha = 0.2$)	0.564	0.593	0.969
Synthetic($\alpha = 0.3$)	0.665	0.673	0.968
Synthetic($\alpha = 0.4$)	0.806	0.816	0.987
Synthetic($\alpha = 0.5$)	0.867	0.903	0.994
Synthetic($\alpha = 0.6$)	0.912	0.964	0.993
AAV-MNIST($\sigma = 2.0$)	0.328	0.334	0.644
AAV-MNIST($\sigma = 1.5$)	0.400	0.423	0.770
AAV-MNIST($\sigma = 1.0$)	0.517	0.567	0.875
AAV-MNIST($\sigma = 0.5$)	0.749	0.780	0.971
AAV-MNIST($\sigma = 0.0$)	0.910	0.955	0.997
AVE	0.301	0.714	0.764
Kinetics-Sounds	0.220	0.570	0.639

Table 7: The uni-modal and multi-modal accuracy of datasets. For synthetic data, X refers to the modality with input dimension $= 100$, and Z refers to the modality with input dimension $= 200$. For AAV-MNIST, Kinetics-Sounds, and AVE datasets, X refers to the audio, and Z refers to the visual input. The uni-modal-X means the accuracy of a classifier trained on only the X-input. Similarly, we report the uni-modal-Z and multi-modal accuracy. For AAV-MNIST, we use the models pre-trained on AAV-MNIST ($\sigma = 0$). For AVE and Kinetics-Sounds dataset, we use the models pre-trained on two modalities respectively.

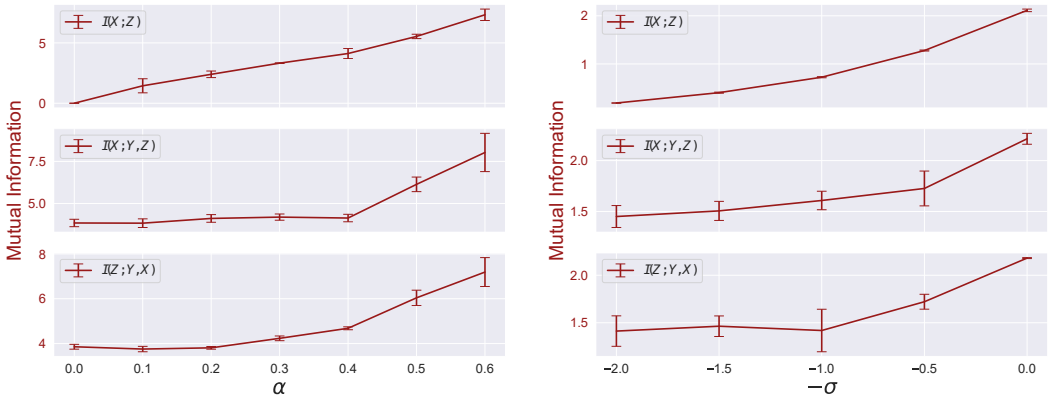

Figure 4: Line plots showing the mean value of estimated mutual information terms with error bars (standard variances) on the synthetic dataset (left) and AAV-MNIST dataset (right). The mean values and standard variances are calculated from three independent repeated experiments. The x-axis is the parameter used in data generation. For synthetic dataset, we plot α. For AAV-MNIST, we plot $-\sigma$ for unity. For synthetic data, X refers to the modality with input dimension $= 100$, and Z refers to the modality with input dimension $= 200$. For AAV-MNIST, X refers to the audio, and Z refers to the visual input.

larger than that of other two terms, leading to the decrease in our metric. The variance of mutual information terms is low except for the cases where $\alpha > 0.5$ in synthetic dataset. In the cases where $\alpha > 0.5$, the three mutual information terms all have large values and correspondingly high variance, resulting in the high variance of the metric which is close to zero. However, from our results of real-world datasets, we observed that the value of three terms are small and thus similar to the cases where $\alpha \leq 0.5$. Hence, our metric is still indicative and effective for real-world datasets. In addition, we provide the numerical results of mutual information terms and their standard variance in the table 8 and table 9 for reference.

α	mean of I(X;Z)	stdvar of I(X;Z)	mean of I(X;Y,Z)	stdvar of I(X;Y,Z)	mean of I(Z;Y,X)	stdvar of I(Z;Y,X)	mean of metric	stdvar of metric
0.0	0.0000	0.0000	3.8451	0.2196	3.8578	0.1079	11.1130	0.4509
0.1	1.4447	0.5831	3.8342	0.2563	3.7549	0.1201	6.7805	1.2357
0.2	2.3966	0.2724	4.1149	0.2271	3.8096	0.0599	4.5174	0.5894
0.3	3.3212	0.0289	4.1939	0.1851	4.2337	0.1051	2.5754	0.1991
0.4	4.1203	0.4197	4.1386	0.2219	4.6818	0.0652	0.8365	1.4316
0.5	5.5372	0.1813	6.1381	0.4324	6.0411	0.3402	1.5937	1.5846
0.6	7.3310	0.4721	8.0298	1.1352	7.1899	0.6451	0.8045	2.1281

Table 8: The numerical results of mutual information terms and their standard variance on the synthetic dataset with different hyperparameters. The mean values and standard variances are calculated from three independent repeated experiments. X refers to the modality with input dimension $= 100$, and Z refers to the modality with input dimension $= 200$.

σ	mean of I(X;Z)	stdvar of I(X;Z)	mean of I(X;Y,Z)	stdvar of I(X;Y,Z)	mean of I(Z;Y,X)	stdvar of I(Z;Y,X)	mean of metric	stdvar of metric
2.0	0.1889	0.0043	1.4512	0.1076	1.4122	0.1604	1.0807	0.1128
1.5	0.4038	0.0112	1.5063	0.0936	1.4635	0.1083	0.9401	0.0773
1.0	0.7248	0.0115	1.6086	0.0908	1.4183	0.2237	0.6858	0.1296
0.5	1.2790	0.0104	1.7265	0.1709	1.7212	0.0784	0.3868	0.1163
0.0	2.1130	0.0266	2.2140	0.0535	2.1807	0.0051	0.0733	0.0262

Table 9: The numerical results of mutual information terms and their standard variance on the AAV-MNIST dataset with different hyperparameters. The mean values and standard variances are calculated from three independent repeated experiments. X refers to the audio, and Z refers to the visual input.

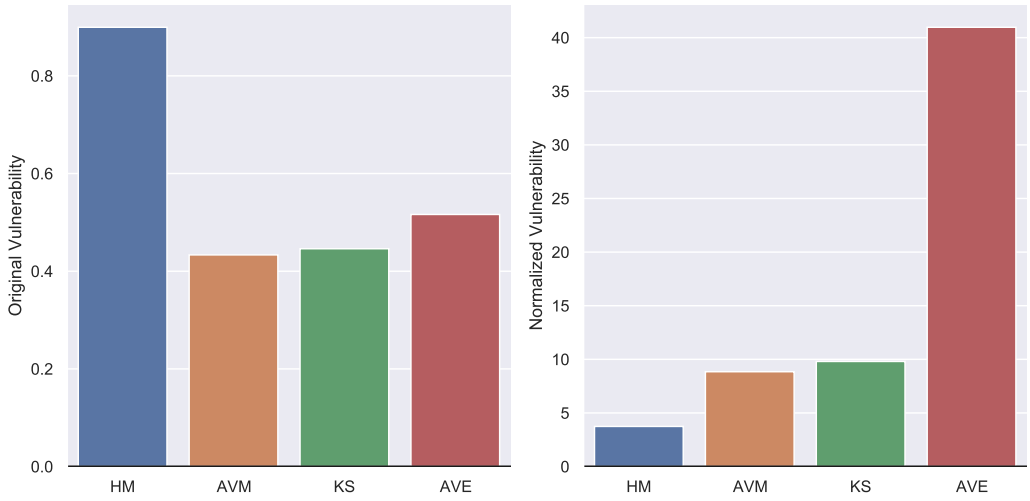

Figure 5: Original and the normalized vulnerability in the modality missing case of four datasets: Hateful-Meme (HM), AV-MNIST (AVM), Kinetics-Sounds (KS), and AVE. All robustness values and metrics used for calculation can be found in table 1.

C.3 Applying the Metric

From our analysis, complementarity is an essential factor in multi-modal robustness. For any model trained on a highly complementary multi-modal dataset, it is impossible to achieve the best accuracy when one modality is corrupted, especially missing, in the test. This is proved in section 3. In other words, no model can avoid the impact of noisy or missing modality. Hence, if we would like to eliminate the influence of modality complementarity and focus on other factors, e.g., adversarial training and model structure, we could normalize the value by our metric.

This subsection demonstrates an example of a simple normalization in the modality missing case. Here, we define the missing vulnerability as (1 - missing robustness). Then we divide the vulnerability of four multi-modal datasets by their metric. The original and the normalized vulnerability are shown in Figure 5. On the left, the model trained on Hateful-Meme seems more vulnerable than others. However, considering that the Hateful-Meme is more complementary according to its high metric, we should attribute this vulnerability much to the intrinsic property of the dataset. At right, the normalized vulnerability of the model trained on Hateful-Meme is lower than others. Therefore, our metric helps the attribution of vulnerability.

D Limitations and Future Work

The first limitation of our work is that our computation of the proposed metric is an optimization process, which leads to a small computational overhead and a concern on error margin.

Second, our study does not reveal all factors of multi-modal robustness. In section 5, we suggest a possible factor, the number of robust features in each modality, which deserves further work to confirm. Other factors that affect the robustness of multi-modal models include the model architecture (e.g., CNN, ViT, and various fusion methods) and the training pipeline (e.g., pre-training, adversarial training), which are used to achieve better performance or robustness. Apart from architecture and pipeline, factors like the anti-noise ability of each modality also matter (e.g., maybe images are more robust against noise than audios). However, we could study their property using uni-modal settings and tools. This paper focuses on modal complementarity, which is brought by the interconnection of modalities. So it distinguishes multi-modal robustness study from uni-modal robustness study, and this is the reason why we pay so much attention to this factor.

There are several possible directions for future work. First, to lower the estimation variance for our metric, the estimator can be improved by substituting MINE with more efficient and accurate estimators or adjusting the network architecture of the estimator.

Second, other factors that affect multi-modal robustness can be pointed out and analyzed. For example, how do different fusions affect the multi-modal robustness, and what are their mechanisms? Can we distinguish the impacts of different factors on multi-modal robustness and diagnose the problem of a given vulnerable multi-modal model? How do we combine these factors to achieve maximal robustness?

Finally, more fine-grained analysis on multi-modal attack and robustness can be conducted. For instance, do the adversarial attacks damage the complementary part more or the shared part more? Conversely, do the defence methods in uni-modal learning work for multi-modal learning, and which part do they benefit? Can we extract the robust features and non-robust features in multi-modal learning like what researchers have done in uni-modal learning? How do we improve the multi-modal robustness based on our knowledge of the interconnection of modalities (e.g. modal complementarity) of a specific task?