Understanding and Measuring Robustness of Multimodal Learning

Nishant Vishwamitra*, Hongxin Hu*, Ziming Zhao*, Long Cheng†, and Feng Luo†

*Computer Science and Engineering, University at Buffalo
†School of Computing, Clemson University
*{nvishwam, hongxinh, zimingzh}@buffalo.edu, †{lcheng2, luofeng}@clemson.edu

Abstract

The modern digital world is increasingly becoming multimodal. Although multimodal learning has recently revolutionized the state-of-the-art performance in multimodal tasks, relatively little is known about the robustness of multimodal learning in an adversarial setting. In this paper, we introduce a comprehensive measurement of the adversarial robustness of multimodal learning by focusing on the *fusion* of input modalities in multimodal models, via a framework called MUROAN (MUltimodal RObustness ANalyzer). We first present a unified view of multimodal models in MUROAN and identify the fusion mechanism of multimodal models as a key vulnerability. We then introduce a new type of multimodal adversarial attacks called decoupling attack in MUROAN that aims to compromise multimodal models by decoupling their fused modalities. We leverage the decoupling attack of MUROAN to measure several state-of-the-art multimodal models and find that the multimodal fusion mechanism in all these models is vulnerable to decoupling attacks. We especially demonstrate that, in the worst case, the decoupling attack of MUROAN achieves an attack success rate of 100% by decoupling just 1.16% of the input space. Finally, we show that traditional adversarial training is insufficient to improve the robustness of multimodal models with respect to decoupling attacks. We hope our findings encourage researchers to pursue improving the robustness of multimodal learning.

Introduction

Multimodal learning has been gradually gaining focus of the research community over the past few years. The approaches for multimodal learning have come a long way from simple models re-purposed for multimodal tasks, to deep learning-based models that are specifically designed for multimodal tasks (referred to as Deep Multimodal Models or DMMs throughout this paper). For example, recent advances in this field have led to several state-of-the-art DMMs, such as ViLBERT (Lu et al. 2019), VisualBERT (Li et al. 2019), MMBT (Kiela et al. 2019), and Pythia (Jiang et al. 2018), while also engendering the collection of several multimodal datasets, such as Hateful Memes (Kiela et al. 2020), Visual Question Answering (VQA) (Goyal et al. 2017), and Visual Commonsense Reasoning (VCR) (Zellers et al. 2019). Due to the success of these DMMs on standard benchmarks, there have been many encouraging attempts to adopt them to real-world and safety-critical scenarios, such as assistance to blind people (Gurari et al. 2018), hate-speech moderation on social media (Kiela et al. 2020), as well as emerging domains, such as Google MUM search (goo 2021). However, in spite of the recent advances, the robustness of DMMs remains poorly understood.

A significant difference between DMMs and their unimodal counterparts is the *fusion* mechanism in DMMs. This fusion mechanism fuses multiple input modalities to learn their joint representation, which is then processed by several fully connected layers to predict classification scores depending on the nature of the corresponding downstream tasks. Different DMMs (Lu et al. 2019; Kiela et al. 2019; Li et al. 2019; Jiang et al. 2018) employ different strategies to learn strong fusion embeddings of their input modalities. This fusion mechanism presents new challenges towards studying the adversarial robustness of these models.

Recently, several unimodal adversarial attacks for deep unimodal models have been formulated to study their robustness. For example, unimodal adversarial images (Szegedy et al. 2013; Madry et al. 2017; Papernot et al. 2016; Wicker, Huang, and Kwiatkowska 2018; Carlini and Wagner 2017; Wang et al. 2020) and unimodal adversarial text (Alzantot et al. 2018; Li et al. 2018; Jin et al. 2019; Alzantot et al. 2018; Ren et al. 2019) have been widely studied, which have exposed numerous vulnerabilities in the deep unimodal models. However, these attacks cannot be directly employed to study the robustness of their deep multimodal counterparts. First, since these attacks can only be applied to single modalities, they do not affect the fusion mechanism that is fundamental to DMMs. Second, since DMMs combine several different types of modalities (e.g. image, text, speech, etc.), a single unimodal attack cannot be used for all those modalities. We note that formulating comprehensive methods to study the robustness of DMMs is of utmost importance to adopting them in real-world systems, such as VQA.

To address these challenges, in this work, we first highlight how multimodal adversarial attacks based on decoupling the input modalities in DMMs can easily compromise these models. Then, we introduce a framework called MUROAN to study the robustness of DMMs based on decoupling of modalities, thereby revealing vulnerabilities in the fusion mechanism of existing DMMs. MUROAN uses a unified view of DMMs to expose its key vulnerability. Then,

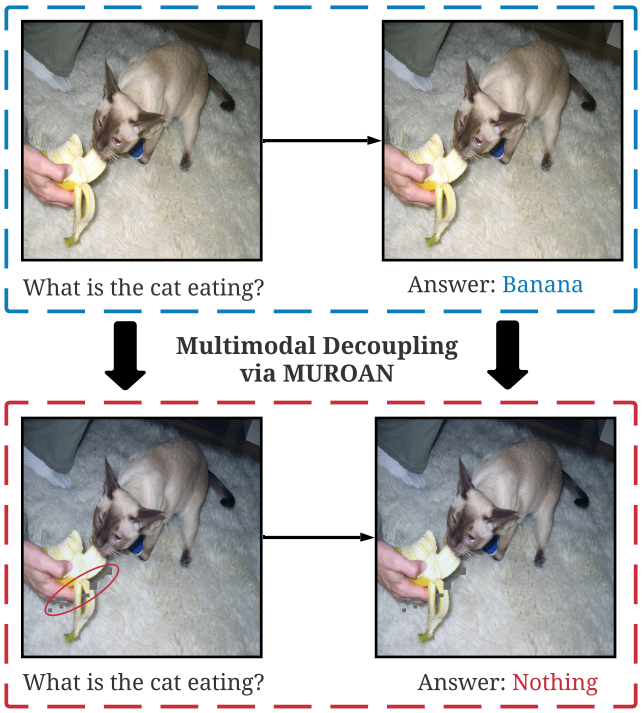


Figure 1: By decoupling the input modalities through removal of a few datapoints in the image via MUROAN framework, the multimodal model predicts a wrong answer class: *Nothing*, indicating that decoupling attack can easily compromise multimodal models.

we introduce a new type of adversarial attacks called decoupling attack in MUROAN, wherein the objective of its attack algorithm is to decouple the input modalities of multimodal models to induce a misclassification. As depicted in Figure 1, a decoupling of the image and text modalities through occlusion of a few datapoints in the image induces a misclassification. In addition, we leverage the MUROAN framework to measure several state-of-the-art DMMs. We find that the seemingly straightforward decoupling attack of MUROAN is in fact highly effective in compromising DMMs.

Our contributions in this work are as follows.

- We present a unified view of DMMs to explore their vulnerabilities, and identify the fusion mechanism of these models as a critical component for their robustness analysis.
- We propose a novel framework called MUROAN that consists of the unified view to exploit the fusion mechanism and a decoupling attack algorithm for comprehensively studying the adversarial robustness of DMMs. MUROAN directly focuses on the fusion mechanism of DMMs by decoupling the input modalities that are fused together.
- We use MUROAN for a comprehensive robustness analysis of state-of-the-art DMMs under several dataset and model settings. Our experiments show that, in the worst case, the decoupling attack in MUROAN can achieve an attack success rate of 100% after decoupling of 1.16% of input modalities of DMMs, while the unimodal adversarial attacks overestimate the robustness of DMMs.

We are open-sourcing our code to encourage research in training DMMs robust to decoupling attacks: http://github.com/SecurityAndPrivacyResearch/mda.

Background

In the following, we give an overview of the field of multimodal learning as well as the state-of-the-art unimodal adversarial attacks used for the robustness analysis of unimodal models.

Multimodal Learning

The renewed interest in multimodal learning can be attributed to more powerful models (Devlin et al. 2018; Vaswani et al. 2017) that can learn strong fusion of input modalities and the availability of several multimodal datasets (Goyal et al. 2017; Zellers et al. 2019; Kiela et al. 2020). These models and datasets have resulted in DMMs achieving impressive results on standard benchmarks. Much of the DMMs that have achieved impressive performances can be categorized under the following categories.

Traditional Fusion-based Models. Several DMMs have attempted to address how to effectively combine multimodal information (Baltrušaitis, Ahuja, and Morency 2018; Bruni, Tran, and Baroni 2014; Lazaridou, Pham, and Baroni 2015). Feature concatenation is one of the most preferred fusion techniques in these models, while some of the models use other feature fusion techniques such as element-wise product. Since these models showed impressive performances on several multimodal benchmarks, they are considered strong baselines for many multimodal tasks.

Transformer-based Fusion Models. Recently, the BERT model (Devlin et al. 2018), a type of transformer (Vaswani et al. 2017), has been shown to achieve state-of-the-art performance (Kiela et al. 2019; Li et al. 2019; Su et al. 2019) on multimodal benchmarks, by learning the interaction between the input modalities via self-attention over many different layers. For example the MMBT (Kiela et al. 2019) model fuses image embeddings in the form of pooled filter maps from a ResNet model and word tokens as two segments of BERT (Devlin et al. 2018). Similarly, the VL-BERT (Su et al. 2019) model fuses regions of interest (ROIs) of an image with word tokens as two segments of BERT. As shown by these works, the transformer based DMMs outperform their unimodal counterparts in multimodal tasks by quite a large margin.

Unimodal Adversarial Attacks

The discovery of unimodal adversarial attacks has engendered active research in the safety and robustness of unimodal deep learning models. In this section, we discuss important unimodal adversarial attacks on images and text.

Unimodal Adversarial Image. A large body of adversarial attacks have been introduced in recent times that mainly

focus towards robustness analysis of computer vision models. For example, several works, such as fast-gradient attacks (Goodfellow, Shlens, and Szegedy 2014; Liu et al. 2016), optimization-based methods (Szegedy et al. 2013; Carlini and Wagner 2017), and other such methods (Papernot et al. 2016; Nguyen, Yosinski, and Clune 2015), have been proposed successfully. Recently, ensemble based attacks (Croce and Hein 2020) have been show to more deeply reveal vulnerabilities in unimodal models. Furthermore, alarmingly critical real-world attacks such as adversarial patches (Brown et al. 2017) have been introduced recently, which cast serious questions on the safety of these vision models.

Unimodal Adversarial Text. Recently, some works have focused on unimodal adversarial text to study robustness of Natural Language Processing (NLP) models. While earlier works (Li et al. 2018; Gao et al. 2018; Eger et al. 2019) effectively employed character level perturbations to perform adversarial attacks, more recent works have found word replacement strategies (Jin et al. 2019; Alzantot et al. 2018; Ren et al. 2019) to be largely effective in compromising these models. Recent works (Iyyer et al. 2018; Zhao, Dua, and Singh 2017; Ribeiro, Singh, and Guestrin 2018) have also demonstrated how sentences can be merely reconfigured to pose serious adversarial threats.

Recently, some studies have emerged that discuss adversarial attacks on DMMs (Tian and Xu 2021; Li et al. 2020). However, these studies do not focus on exploring the vulnerabilities of the fusion mechanism to adversarial attacks. In this work, we specifically focus on comprehensively studying the adversarial robustness of DMMs via a type of multimodal adversarial attack called decoupling attack, that focuses on decoupling the input modalities of a DMM to compromise the fusion mechanism of these DMMs.

Decoupling Input Modalities

Our primary objective in this section is to demonstrate how easily decoupling of input modalities can compromise DMMs. To this end, we performed a preliminary experiment for comparing the effect of decoupling attacks on DMMs against traditional unimodal adversarial attacks.

We randomly selected 100 samples from the VQA dataset (Antol et al. 2015) and the pretrained Pythia DMM (Jiang et al. 2018) to conduct our preliminary study. Several previous unimodal attacks (Goodfellow, Shlens, and Szegedy 2014; Madry et al. 2017; Kurakin, Goodfellow, and Bengio 2016; Moosavi-Dezfooli, Fawzi, and Frossard 2016; Papernot et al. 2016; Xie et al. 2019; Dong et al. 2018) have revealed the nature of different vulnerabilities in traditional unimodal model. In this experiment, we use the state-of-the-art attack called PGD attack (Madry et al. 2017) as the traditional, unimodal adversarial attacks. Next, we manually studied the 100 samples and occluded datapoints that we considered as participating in the fusion mechanism. Our objective from this step was to manually decouple the DMMs to study whether decoupling could be considered as an effective means to create adversarial attacks that can be comparable against strong and popular unimodal attacks in terms of their effectiveness in fooling the DMM.

We found that just manual decoupling was able to effectively fool 50% of the samples considered in the experiment. But more importantly, we found that on average, manual decoupling only affected 7.25% of the datapoints in the image of each sample. On the other hand, we found that although the PGD attack was quite effective in compromising the DMM with close to 100% success rate, 96.47% of the datapoints on average were affected by PGD. What this experiment shows is that unimodal adversarial attacks are not able to identify the optimum datapoints to perturb. Thus, unimodal attacks are not sufficiently suitable for studying robustness of DMMs. Furthermore, since unimodal attacks do not seem to take the fusion mechanism into consideration, they do not reveal the vulnerabilities specific to DMMs. In the sections that follow, we show how MUROAN decoupling attack algorithm can optimally find the exact datapoints involved in fusion, so that the adversarial robustness through decoupling can be studied.

Robustness Analysis

In this section, we discuss our approach for the robustness analysis of DMMs via MUROAN framework. In this regard, we first discuss a unified view of DMMs to explore the vulnerabilities of the fusion mechanism of DMMs, and then introduce our algorithm to decouple the fused modalities of DMMs. The overview of our approach is depicted in Figure 2.

Unified View of Deep Multimodal Models

We consider a DMM $D : X \rightarrow Y$ to be a function that maps a domain X to a co-domain Y. An input is a set of vectors of different modalities $x = \{x_0^1 \dots x_n^1, x_0^2 \dots x_m^2, \dots\}$ (Figure 2, Step (a)). We consider Y to be the set of possible classes for a multimodal input $x \in X$. The output of the DMM for a multimodal input x is considered to be $D(x) = y$, for some $y \in Y$. We denote the confidence of the DMM for a multimodal classification probability on input x and class y as $D_y(x)$. Lastly, we denote the cardinality of a set as $|\cdot|$, which represents the number of elements in the set.

Although DMMs have several different architectural configurations, we need a unified view (or representation) of them for a uniform vulnerability analysis of all these different multimodal architectures. To achieve this, we unify these different architectural approaches into a single view, in which we consider a DMM as a generator of the fusion embedding of multiple input modalities (Figure 2, Step (b)), followed by several fully connected layers that are specific for downstream tasks. In other words, we break down a DMM into two functions: the first generates a latent representation (i.e., the fusion embedding) of the multimodal inputs and the second performs classification based on the fusion embedding. We consider the fusion embedding of a multimodal input x as $Z(x) = z$, where z is the d-dimensional fusion embedding vector. Next, we consider $y = M(z)$ to represent classification based on the fusion embedding from fully connected layers that are specific to downstream tasks. Therefore, the original DMM is broken down into two functions, represented as $M(Z(x))$. We

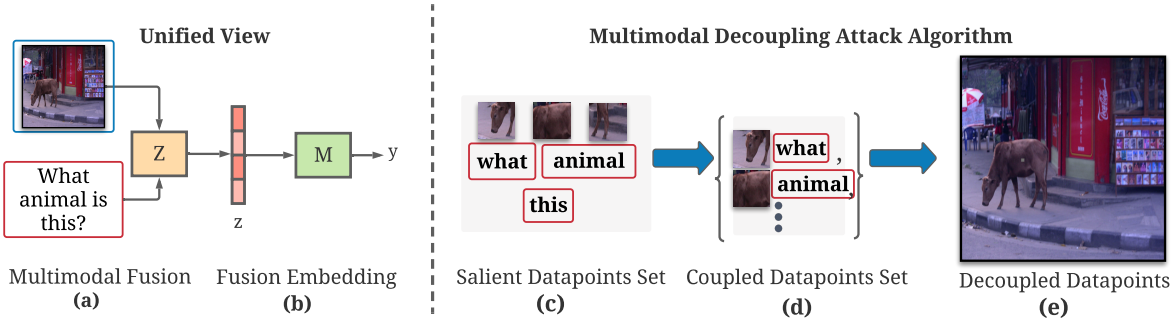


Figure 2: Overview of our approach.

further discuss this process for two typical DMM architectures: traditional architectures and transformer-based architectures.

Traditional Multimodal Architectures. Traditional DMM architectures are composed of separate neural networks that are specific to each input modality, whose outputs are combined using fusion techniques such as element-wise multiplication, addition or concatenation. For example, the Pythia (Jiang et al. 2018) architecture is composed of a convolutional neural network that learns the embedding of the image modality, and a recurrent network that learns the embedding of the text modality, which are then combined using element-wise multiplication. This combination represents the fusion embedding.

Transformer-based Multimodal Architectures. These architectures use the transformer (Vaswani et al. 2017) for learning a strong fusion embedding of the input modalities. The input modalities are first converted into embeddings, which are then combined using the transformer, which performs several self-attentions across many layers. The first token embedding then constitutes the fusion embedding, which is subsequently processed by fully connected layers for classification.

MUROAN Framework

We note that the traditional methods of adversarial attacks are not suitable for DMMs for two specific reasons. First, most key methods of crafting adversarial attacks use either the l_∞ or l_2 norm. Optimization with respect to these kinds of manipulations induces a perturbation in all (or almost all) of the datapoints of an input modality by a small value $\pm\epsilon$. This is not suitable in case of multimodal inputs because different modalities have different compositions, and not all modalities support this type of manipulation. For example, image-based inputs are continuous and thus suitable for such manipulations, but text-based inputs are discrete, thus not suitable for such manipulations. Furthermore, for DMMs, such adversarial manipulations are not suitable for robustness analysis processes since the core weaknesses of these models should be examined in the fusion mechanism of these models, which is not achieved by these manipulations. Since we are interested in studying the effect of decoupling fused modalities, we employ l_0-norm optimization attack algorithm, wherein an l_0-norm attack optimizes for the number of changes made to the inputs for a successful decoupling attack.

Removal of salient datapoints from inputs has been shown to be an important factor for considering the robustness and safety of a decision model (Mathias et al. 2013; Wicker and Kwiatkowska 2019; Noh et al. 2018). However, the key difference between the traditional unimdoal domains and the multimodal domain is that such datapoints are in fact parts of separate modalities that are coupled together by the multimodal fusion mechanism. Thus, it is imperative to study the cases, in which some parts of the input modalities are removed, so as to render this fusion as unsuccessful.

For a multimodal input x, we consider coupled datapoints as some $x' \subset x$. Our objective is to find the minimum subset via the following optimization.

$$\underset{x' \subset x}{\arg\min}(|x| - |x'|) \ni D(x) \neq D(x') \tag{1}$$

However, it is impractical to solve the optimization in Equation 1, due to a large number of such datapoints in the multimodal input space. Thus, to solve this optimization, we use the notion of the fusion embedding to compute a salient points set first, S_n (Figure 2, Step (c)). Such sets of critical or salient points have been previously utilized to inspect deep learning-based models (Qi et al. 2017). We use the salient datapoints set to study the weaknesses of DMMs, by defining it as follows.

$$S_n^x = \{x_i \in x \mid Z(x/x_i) \neq Z(x)\} \tag{2}$$

In Equation 2, the salient datapoints set contains those datapoints that affect the fusion embedding upon removal (where x/x_i denotes removal of a datapoint). For example in the transformer-based DMMs, a datapoint $x_i \in S_n^x$ if $\forall i \neq j, z_i \geq z_j$ due to the transformer pooling layer. Next, we find the set of coupled datapoints (Figure 2, Step (d)) from the salient datapoints set, by computing permutations of all datapoints of a maximum size equal to the size of the salient datapoints set (denoted by $\prod$ in Equation 3).

$$C_n^x = \prod_{i=1}^{||S_n^x||} \{s_i\} \tag{3}$$

Algorithm 1: MUROAN Decoupling Attack Algorithm

Input: x, y, D, Θ, f, **maxitr**
Output: x'

Initialization: $x' \leftarrow x$
while *f(D, x, x′, y)* or **maxitr do**
  $S_n^x \leftarrow \text{GetSalientSet}(D, x')$
  $C_n^x \leftarrow \text{GetCoupledSet}(S_n^x)$
  for $x_i \in C_n^x$ **do**
    if $D(x') \neq y$ **then**
      break
    end
    $x' \leftarrow x'/x_i$
    if $D(x') \neq y$ and $f(D, x, x', y) \neq$ True **then**
      $x \leftarrow x$
    end
  end
end
return x'

Now that we have computed the coupled datapoints set, we propose the MUROAN Decoupling Attack Algorithm (Algorithm 1) to iteratively refine the decoupling attack. In our algorithm, first the salient datapoints set is computed based on the process described in Equation 2. Then, the GetCoupledSet procedure is called, which performs two functions. First, the coupled datapoints are computed as described in Equation 3. Then, they are ordered based on the size of the datapoints, so as to satisfy Equation 1. We encode the termination of our algorithm as a boolean function f, to support multiple adversarial requirements. For example, adversarial requirements for crafting untargeted attacks ($D(x') \neq y$) or targeted attacks ($D(x') = y'$) can be supported (Figure 2, Step (e)). Lastly, we propose the following theorem to use our decoupling attack algorithm as a robustness verification technique to find adversarial examples in DMMs if one exists.

Theorem 1. *For a multimodal model D that satisfies our unified view and a given multimodal input x, the MUROAN decoupling attack algorithm will find the optimum adversarial example that satisfies Equation 1.*

Proof. If an adversarial example exists for input x, it can be found by an exhaustive search of the input space. The GetCoupledSet function returns all possible permutations of the coupled datapoints and the f function and **maxitr** can be set such that the algorithm does not terminate until a satisfactory adversarial permutation is found. Furthermore, since the permutations in the coupled datapoints set are ordered, thus, a permutation that is found by our algorithm to be adversarial is minimal.

Experiments

In our evaluation, we use MUROAN to analyze the robustness of state-of-the-art DMMs trained on popular multimodal datasets to show how decoupling attack can easily compromise these models, thereby enabling us to understand their robustness. We also consider some unimodal adversarial attack baselines in our evaluation only to show how easily decoupling attack can compromise DMMs. Our objective is not to make a direct comparison of our approach against these existing attacks, but to highlight how decoupling of input modalities can be easily used to attack the fusion mechanism of DMMs. Our findings highlight the need for rigorous safety analysis of DMMs against decoupling attacks, and lay down important groundwork for their deployment in real-world applications. We first summarize the DMMs, datasets, and unimodal adversarial baselines that are used in our experiments.

Deep Multimodal Models

- **Pythia**. The Pythia (Jiang et al. 2018) is a state-of-the-art model in the VQA challenge task. This models is composed of a convolutional network to compute an image embedding and a recurrent network to compute a sentence embedding, which are fused using element-wise multiplication.
- **Late Fusion**. We consider the late-fusion architecture based DMM in (Antol et al. 2015) as a strong baseline model. In this model, image embeddings from a convolutional neural network and text embeddings from a recurrent network are fused using element-wise sum, and then the fusion embedding is processed through multiple classification layers to generate a probability score.
- **MMBT**. The MMBT model (Kiela et al. 2019) is a state-of-the-art DMM that utilizes the BERT (Devlin et al. 2018) to learn multimodal embeddings by the implicit alignment of image and text features with the self-attention mechanism of transformers (Vaswani et al. 2017), for a wide range of visual-linguistic tasks. The query vector of this model, which is treated as the fusion embedding, is processed through a classifier head for downstream tasks.

Multimodal Datasets

- **Hateful Memes**. The Hateful Memes (Kiela et al. 2020) dataset consists of image and text pairs pertaining to hateful memes, a recent phenomenon that poses a serious threat societal threat in today's day and age. The objective is classification into two categories: "hateful" or "non-hateful".
- **Visual Question Answering (VQA)**. The VQA dataset (Antol et al. 2015) consists of images with multiple associate natural language questions. Each image and question pair expects a list of answers. The objective is to predict the best answer from the list of answers for each image-question pair.

Unimodal Adversarial Baselines

- **CW Attack**. We use the Carlini and Wagner (Carlini and Wagner 2017) attack algorithm as baseline for unimodal adversarial images for image-based modality.
- **Genetic Attack**. We use the Genetic Attack (Alzantot et al. 2018) algorithm (referred to as "Genetic" in this

What is the man pulling?
Answer: Nothing
(a)

Safe sex isis style
(b)

White power!
(c)

Figure 3: Three samples depict three types of minimum coupled datapoints in the VQA and Hateful Memes dataset. In sample (a), the minimum coupled datapoints are in the image only (indicated by red circles), and it is enough to only make changes to a those datapoints to decouple the sample. In sample (b), the minimum coupled datapoints are in the text only (indicated by red font), it is enough to make changes to the text only to decouple the sample. In sample (c), the coupled datapoints consist of both image and text, therefore both need to be changed to decouple this sample.

paper) as baseline for unimodal adversarial text for text-based modality.

Other Implementation Details

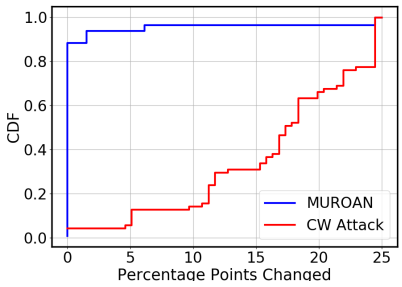
We have implemented our attack using the PyTorch (Paszke et al. 2019) library. For the VQA dataset we used 1000 samples and for Hateful Memes dataset, we used 250 samples to conduct our experiments. We used pretrained models (pre 2021) published by the original authors for all the DMMs that we have evaluated in our experiments. In the MUROAN decoupling attack algorithm, we used a maximum iteration limit of 500 epochs, post which we report the attack as unsuccessful. We ran our experiments on a single NVIDIA V100 GPU enabled eight core machine.

Robustness Analysis

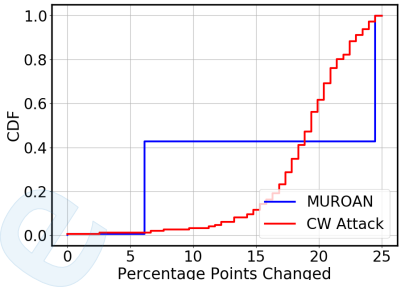
In this section, we used our framework to analyze the robustness of state-of-the-art DMMs under various attack conditions to show that the robustness of these DMMs are largely overestimated.

We studied the percentage of points changed by MUROAN decoupling attack algorithm in comparison with the CW attack for a successful misclassification. We used the same cutoff of 500 epochs for both the algorithms in all the tests, post which we reported a failure. We have depicted the results of this experiment in Figure 4.

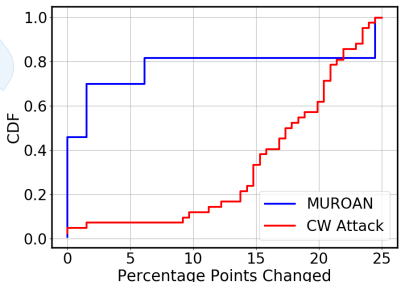
Figure 4 depicts the CDF of the average percentage of datapoints changed in both the attacks under consideration. We found that the unimodal adversarial images (i.e., the CW attack) vastly overestimated the robustness of all the three DMMs. For the Pythia-VQA, it was observed that the CW attack changed 93.99% of the input datapoints, whereas MUROAN decoupling attack algorithm changed 1.16% of input datapoints. This difference of a large margin showed that the baseline unimodal adversarial images vastly overestimated the robustness of models for the VQA task. This finding may have important implications on using VQA in



(a) Pythia-VQA



(b) Late Fusion-Hateful Memes



(c) MMBT-Hateful Memes

Figure 4: CDF of percentage of datapoints changed.

real-world applications, such as visual question answering for the blind (Gurari et al. 2018). Next, we discuss another important application domain, namely Hateful Memes. For the Late Fusion- Hateful Memes model, it was again observed that the CW attack changed 99.86% of datapoints, whereas MUROAN decoupling attack algorithm changed an average of 16.93% of input datapoints, a significant difference. For the MMBT-Hateful Memes model, it was observed that the CW attack changed 94.92% of datapoints, whereas MUROAN decoupling attack algorithm changed an average of 5.73% of input datapoints. In both cases, the unimodal adversarial images overestimated the robustness of the DMMs trained for the Hateful Memes task.

Next, we compared the Attack Success Rate (ASR) of MUROAN decoupling attack algorithm with respect to the unimodal adversarial images and text baselines, namely the CW (Carlini and Wagner 2017) attack and the Genetic (Alzantot et al. 2018) attack respectively. The results of this experiment have been depicted in Table 2. We first

Model-Dataset	Average Points Changed - MUROAN	Average Points Changed - CW
Pythia-VQA	**1.16%**	93.99%
Late Fusion-Hateful Memes	**16.93%**	99.86%
MMBT-Hateful Memes	**5.73%**	94.92%

Table 1: Comparison of Average Percentage Points Affected by MUROAN and CW attack.

Model-Dataset	ASR-MUROAN	ASR-CW	ASR-Genetic	ASR-CW+Genetic
Pythia-VQA	**100%**	79.77%	49.30%	86.45%
Late Fusion-Hateful Memes	**97.25%**	59.11%	0%	59.11%
MMBT-Hateful Memes	**83.33%**	47.19%	0%	47.19%

Table 2: Comparison of Attack Success Rate (ASR).

discuss the impact of the unimodal adversarial images on the DMMs. In all the three DMMs, we found that the unimodal adversarial images could affect these DMMs. However, they vastly overestimated their robustness in all three cases, when we compared the ASRs of MUROAN decoupling attack algorithm. For the Pythia-VQA model, the CW attack achieved an ASR of 79.77%, although the ASR achieved by MUROAN decoupling attack algorithm was 100%. For the two DMMs for hateful memes (i.e., Late Fusion-Hateful Memes and MMBT-Hateful Memes), a similar observation was made, although the CW attack achieved significantly lower ASR for both DMMs. Next, we took a closer look at the impact of the unimodal adversarial text (i.e., Genetic attack) on the DMMs, in comparison with MUROAN. For the Pythia-VQA, it was observed that the Genetic attack has little effect when compared to MUROAN decoupling attack algorithm, and even to the CW attack, wherein both these attacks outperformed the unimodal adversarial text baseline by a large margin. In case of the hateful memes DMMs (i.e., Late Fusion-Hateful Memes and MMBT-Hateful Memes) this margin was found to be even larger. It was observed that the unimodal adversarial text had no significant effect on the DMMs for hateful memes.

Thus, we observed that the safety and robustness of these DMMs need to be deeply examined, specifically from the perspective of decoupling attacks. In this regard, our experiments indicate that our attack exposes the vulnerabilities in the fusion mechanism of DMMs, and the robustness of this mechanism needs significant improvement, especially if DMMs are to be deployed in real-world systems.

Qualitative Analysis of MUROAN

In this section, we provide a qualitative analysis of the decoupled samples that our the MUROAN decoupling attack algorithm generated. Upon observation of such samples in the two baseline datasets (i.e., VQA and Hateful Memes), we discuss certain aspects of the nature of decoupling pertaining to our observations. In Figure 3 [1], we depict three samples from our robustness analysis experiments. Figure 3 (a) is from the VQA dataset, and Figures 3 (b) and (c) are from the Hateful Memes dataset. These three samples represent the three levels of decoupling we observed in our experiments. In Figure 3 (a), the minimum coupled datapoints were found in the image only, therefore it is sufficient to decouple just the single image modality. In the VQA dataset, since questions are asked about certain parts of an image, this observation is intuitive since it should be sufficient to only affect the relevant parts of the image. In Figure 3 (b), the minimum coupled datapoints were only found in the text modality, since intuitively we cannot see why this sample could be a hateful meme from the image alone. In Figure 3 (c), the minimum coupled datapoints consist of both the image and the text modalities. In this case, both the input modalities need to be affected for decoupling this fusion. Therefore, we note that vulnerabilities in the DMMs are of a very different nature when compared to their unimodal counterparts.

[1]Note: samples (b) and (c) are from the Hateful Memes dataset (Kiela et al. 2020), which some readers may find distressing.

Adversarial Training

Our experiments in Section 15 raise an important question: how can we defend against decoupling attacks? We performed a preliminary experiment to see if adversarial training (Goodfellow, Shlens, and Szegedy 2014), a popular technique to improve adversarial robustness, can be used to reduce the attack success rate. We performed adversarial training using the MMBT model for the hateful memes classification. We generated 247 adversarial examples via MUROAN framework and trained the model on these samples combined with the original dataset from scratch. We observed that the adversarial trained DMM was still vulnerable to newly crafted decoupled samples, despite the model achieving near 100% accuracy classifying adversarial examples included in the training set. These results demonstrate the difficulty in defending against decoupling attacks using traditional adversarial training. We hope these results inspire further work in increasing the robustness of DMMs.

Discussion

In this section, we discuss some limitations, potential negative societal impacts, and some future directions of our work.

In this work, we have focused on DMMs that mainly operate on image and text modalities as inputs. We chose this type of DMMs since it could represent different compositions of inputs (i.e., a continuous input and a discrete input).

Our approach however can be generalized to incorporate any other types of DMMs, considering compositions of other inputs including speech and video modalities.

Our major research objective is to improve the robustness of DMMs by showing their vulnerabilities to decoupling attacks, so that adversarial attacks on real-world multimodal systems can be mitigated. A potential negative societal impact of our work is that it might be used to craft adversarial attacks against DMMs. As future work, we will investigate potential techniques to improve the robustness of DMMs. We hope our findings encourage more researchers to pursue improving the robustness of DMMs.

Conclusion

In conclusion, we have studied the robustness of DMMs against multimodal decoupling attacks that are aimed at compromising the fusion mechanism of DMMs. We have introduced a new framework called MUROAN for studying the robustness of DMMs, which consists of a unified view of the DMMs that exposes the fusion embedding, and an algorithm for decoupling the input modalities. Our experiment regarding adversarial training shows that it does not improve the robustness against our decoupling attacks. MUROAN paves the way for studying the robustness of DMMs via decoupling input modalities in the future.

References

2021. Google MUM: New Technology For Complex Search Queries. https://www.searchenginejournal.com/google-mum-new-technology-for-complex-search-queries/407725/#close. Accessed: 2021-05-19.

2021. Pretrained Models. https://github.com/facebookresearch/mmf/tree/master/projects/hateful_meme. Accessed: 2021-05-26.

Alzantot, M.; Sharma, Y.; Elgohary, A.; Ho, B.-J.; Srivastava, M.; and Chang, K.-W. 2018. Generating natural language adversarial examples. *arXiv preprint arXiv:1804.07998*.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.

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

Brown, T. B.; Mané, D.; Roy, A.; Abadi, M.; and Gilmer, J. 2017. Adversarial patch. *arXiv preprint arXiv:1712.09665*.

Bruni, E.; Tran, N.-K.; and Baroni, M. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49: 1–47.

Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, 39–57. IEEE.

Croce, F.; and Hein, M. 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, 2206–2216. PMLR.

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

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9185–9193.

Eger, S.; Şahin, G. G.; Rücklé, A.; Lee, J.-U.; Schulz, C.; Mesgar, M.; Swarnkar, K.; Simpson, E.; and Gurevych, I. 2019. Text processing like humans do: Visually attacking and shielding NLP systems. *arXiv preprint arXiv:1903.11508*.

Gao, J.; Lanchantin, J.; Soffa, M. L.; and Qi, Y. 2018. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops (SPW)*, 50–56. IEEE.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gurari, D.; Li, Q.; Stangl, A. J.; Guo, A.; Lin, C.; Grauman, K.; Luo, J.; and Bigham, J. P. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3608–3617.

Iyyer, M.; Wieting, J.; Gimpel, K.; and Zettlemoyer, L. 2018. Adversarial example generation with syntactically controlled paraphrase networks. *arXiv preprint arXiv:1804.06059*.

Jiang, Y.; Natarajan, V.; Chen, X.; Rohrbach, M.; Batra, D.; and Parikh, D. 2018. Pythia v0. 1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*.

Jin, D.; Jin, Z.; Tianyi Zhou, J.; and Szolovits, P. 2019. Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment. *arXiv*, arXiv–1907.

Kiela, D.; Bhooshan, S.; Firooz, H.; and Testuggine, D. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*.

Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; and Testuggine, D. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *arXiv preprint arXiv:2005.04790*.

Kullback, S. 1997. *Information theory and statistics*. Courier Corporation.

Kurakin, A.; Goodfellow, I.; and Bengio, S. 2016. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*.

Lazaridou, A.; Pham, N. T.; and Baroni, M. 2015. Combining language and vision with a multimodal skip-gram model. *arXiv preprint arXiv:1501.02598*.

Li, C.; Tang, H.; Deng, C.; Zhan, L.; and Liu, W. 2020. Vulnerability vs. reliability: Disentangled adversarial examples for cross-modal learning. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 421–429.

Li, J.; Ji, S.; Du, T.; Li, B.; and Wang, T. 2018. Textbugger: Generating adversarial text against real-world applications. *arXiv preprint arXiv:1812.05271*.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Liu, Y.; Chen, X.; Liu, C.; and Song, D. 2016. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Mathias, M.; Benenson, R.; Timofte, R.; and Van Gool, L. 2013. Handling occlusions with franken-classifiers. In *Proceedings of the IEEE International Conference on Computer Vision*, 1505–1512.

Moosavi-Dezfooli, S.-M.; Fawzi, A.; and Frossard, P. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2574–2582.

Nguyen, A.; Yosinski, J.; and Clune, J. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 427–436.

Noh, J.; Lee, S.; Kim, B.; and Kim, G. 2018. Improving occlusion and hard negative handling for single-stage pedestrian detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 966–974.

Papernot, N.; McDaniel, P.; Jha, S.; Fredrikson, M.; Celik, Z. B.; and Swami, A. 2016. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, 372–387. IEEE.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32*, 8024–8035. Curran Associates, Inc.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 652–660.

Ren, S.; Deng, Y.; He, K.; and Che, W. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, 1085–1097.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2018. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 856–865.

Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; and Dai, J. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Tian, Y.; and Xu, C. 2021. Can audio-visual integration strengthen robustness under multimodal attacks? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5601–5611.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, B.; Pei, H.; Pan, B.; Chen, Q.; Wang, S.; and Li, B. 2020. T3: Tree-Autoencoder Regularized Adversarial Text Generation for Targeted Attack. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6134–6150.

Wicker, M.; Huang, X.; and Kwiatkowska, M. 2018. Feature-guided black-box safety testing of deep neural networks. In *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, 408–426. Springer.

Wicker, M.; and Kwiatkowska, M. 2019. Robustness of 3d deep learning in an adversarial setting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11767–11775.

Xie, C.; Zhang, Z.; Zhou, Y.; Bai, S.; Wang, J.; Ren, Z.; and Yuille, A. L. 2019. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2730–2739.

Yu, F.; Qin, Z.; Liu, C.; Zhao, L.; Wang, Y.; and Chen, X. 2019. Interpreting and evaluating neural network robustness. *arXiv preprint arXiv:1905.04270*.

Zellers, R.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6720–6731.

Zhao, Z.; Dua, D.; and Singh, S. 2017. Generating natural adversarial examples. *arXiv preprint arXiv:1710.11342*.

Appendix

Additional Qualitative Results

In this section, we provide additional qualitative examples of our attack against the MMBT-Hateful Memes model and the Pythia-VQA model in Figure 5 and Figure 6, respectively. In Section 15, we discussed a few samples from MUROAN from the Hateful Memes dataset. We further discuss more samples from the VQA dataset in addition to some samples from the Hateful Memes dataset in this section.

	who knew that this country is full of ~~white~~ trash	Non-hateful
	islam is a religion of peace stop criticizing my religion	Non-hateful
	told girlfriend that mom is deaf so speak loud and slow told mom that girlfriend is ~~retarded~~	Non-hateful

Figure 5: Additional Samples from the Hateful Memes dataset.

	what object is depicted in this picture?	Original Answer: Stop sign Adverarial answer: House
	~~what~~ sports team is mentioned?	Original Answer: Red sox Adverarial answer: Yes
	what are the ~~color~~ of the lines on the court?	Original Answer: Red and blue Adverarial answer: Red

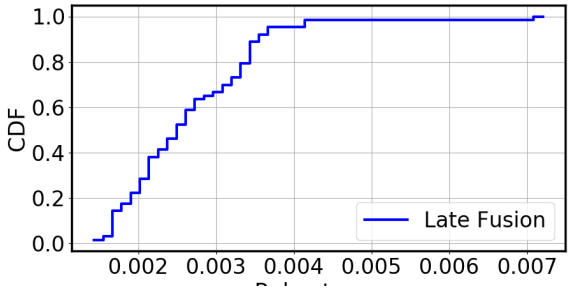
Figure 6: Additional Samples from the VQA dataset.

Figure 5 depicts three samples from the MMBT-Hateful Memes baseline. The first sample depicts the case where only the text is needed to be manipulated to decouple the input modalities in a sample. The second example depicts the case where only a part of the image needs to be manipulated to decouple the modalities in a sample. The third example depicts the case where both image and the text need to be manipulated to decouple the modalities in a sample.

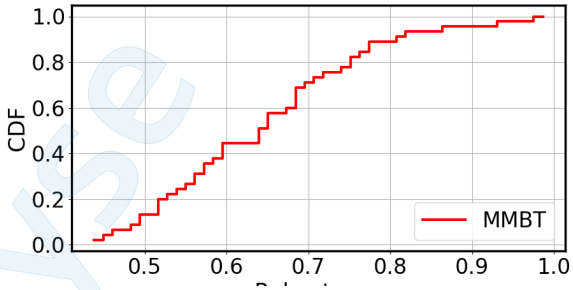
Figure 6 depicts three samples from the Pythia-VQA baseline. In this case, the objective is to fool the DMM so as to output a wrong answer (as opposed to a wrong label in the Hateful Memes case). We observed a similar trend in case of VQA as well, as noted in Section 15. In some cases (such as the first sample and the second sample in Figure 6), it was sufficient to only manipulate one of the input modalities to decouple the input modalities in a sample. In some cases though, both modalities had to be manipulated for decoupling them (such as the third sample in Figure 6).

Quantitative Robustness Analysis of DMMs

We have discussed in Section 15 about how our attack can be used to study the robustness of several DMMs. In this section, we use our attack to study and compare the robustness of two baseline DMMs, Late Fusion and MMBT, discussed in our paper.



(a) Late Fusion Model



(b) MMBT

Figure 7: CDF of robustness of Late Fusion model and MMBT against MUROAN decoupling attack algorithm.

In this experiment, our objective is to compare the two DMMs that are trained for the same task to determine which DMM is more robust against our attack. In this way, we can use MUROAN to additionally compare DMMs in terms of their robustness. We study and compare the robustness of the two DMMs both trained on the Hateful Memes dataset based on the *robustness metric* ψ (Yu et al. 2019). Model robustness is defined as follows.

$$\psi(x) = \frac{1}{\max_{\delta \in set} D_{KL}(P(x), P(x+\delta))} \tag{4}$$

Equation 4 uses the Kullback–Leibler divergence loss (D_{KL}) (Kullback 1997) to depict the divergence between the probability distributions of the original samples and the adversarial samples generated by MUROAN decoupling attack algorithm. In other words, the D_{KL} is higher for a model, for which the adversarial samples are further from the original distribution, indicating stronger robustness. In this experiment, we compared the robustness of the MMBT model to the Late Fusion model, where both DMMs were trained on the same Hateful Memes dataset. The distribution of the robustness the two DMMs as calculated by Equation 4 based on our attack is depicted in Figures 7a and 7b, respectively. We found that the MMBT model is significantly more robust than the Late Fusion model, as can be observed from the Figure 7. The mean robustness of the MMBT model was

found to be $\psi = 0.65$ and the mean robustness of the Late Fusion model was found to be $\psi = 0.003$. The higher robustness of the MMBT model could be attributed to the way the fusion is achieved in this DMM, using the more sophisticated self-attention mechanism of the transformer (Vaswani et al. 2017), while the Late Fusion model uses the element-wise addition. Thus, the robustness metric in this experiment could also indicate the strength of the fusion mechanism. In this way, the robustness of the state-of-the-art DMMs can be quantitatively measured using MUROAN.