CAP-6938
Trustworthy Machine Learning Project

Investigating Adversarial Robustness of Hateful Meme Detection Systems

by

Eftekhar Hossain (ID: 5680934)
Faria Binte Kader (ID: 5662486)

Project Code

Department of Computer Science
University of Central Florida
Orlando, Florida, USA

December, 2024

Members' Contributions

Eftekhar Hossain

I worked with the baseline model training to get the baseline results for evaluation. Next I implemented the white box attacks and trasnfer attack on images and evaluated the models' performances for these attacks.

Faria Binte Kader

I worked with the Black box attacks on images as well as texts and evaluated the different models' performances on the datasets.

Both of us equally contributed to report structuring and writing.

1 Introduction

The proliferation of hate on social media has significantly increased in recent years. Recently, multimodal content such as memes has gained more attraction among users to communicate. However, memes have been widely used to propagate malicious activity due to their implicit humorous characteristics. Earlier social media companies recruit moderators to manually monitor harmful memes and remove them from their media platforms. With the advancement of NLP and computer vision, this task has become easier, as machines can now understand the semantics of hateful memes, thus helping to slow down the spread of such content. However, systems that rely heavily on machine learning rather than empirical principles can sometimes produce unexpected results. Hateful meme detection systems are developed based on multimodal machine learning,g where joint evaluation of multiple modalities (i.e., image and text) is required. A critical difference between unimodal and multimodal systems is that multimodal systems use a fusion mechanism to understand the overall context of memes from both visual and textual information. This mechanism introduces new difficulties in assessing the robustness of these models to adversarial attacks. For example, the adversary can only introduce slight modifications to either an image or text of the meme so that the human can not perceive the changes; however, the model might make mistakes due to these changes. Therefore, these systems are more likely to experience unexpected behavior, which may lead to potential performance issues that can pose significant challenges to their applicability. Considering the potential harm from misuse of these systems, in this work, we study the robustness of existing state-of-the-art hateful memes detection systems against various adversarial attacks. Since a good amount of research (Aggarwal et al., 2023; Evtimov et al., 2020) has been done in this direction for the English language, we study the hateful memes detection systems for the Bangla language in this work. Our main objectives are to *(i)* explore the vulnerability of existing multimodal detection models under various levels of adversarial attacks, *(ii)* examine how these models perform when only image or text of the memes are attacked and *(iii)* investigate the which modality is more vulnerable towards attack. Overall, our contributions to this work are as follows:

- We present a thorough investigation of the adversarial robustness of the Bengali multimodal hateful meme detection systems using two datasets (*BHM* and *MIMOSA*). We explore various image-based adversarial attacks, including white-box and transfer attacks. Our study provides a comprehensive analysis of how these attacks degrade the performance of models existing SOTA models and reveals their vulnerability to subtle and strong adversarial manipulations.

- We compose several black-box adversarial attacks on the memes' image and text. We experiment with noises such as Salt-Pepper, Gaussian, NewsPrint, and Random for image attacks. Additionally, we introduce four types of text-based attacks, including emoji insertion, positive token injection, typos, and word translation, to examine their influence on model performance. Our findings highlight that the current models are more vulnerable to text-based black box attacks than image ones.

2 Literature Review

With the emerging use of LLMs (Large Language Models) and LVLMS (Large Visual Language Models) in various machine learning tasks, it is necessary to check the robustness of these models against various adversarial attacks to see how vulnerable they are. Especially for cross-modal tasks, the models have two different embeddings- images and texts (corresponding captions, queries, or instructions), images, audio, and text, etc., making them more susceptible to adversarial attacks. Thus, many works have started exploring the potential of adversarial attacks for cross-modality. Adversarial attacks on LVLMs or multimodal LLMs can be of several types, white box attacks (usually includes gradient-based attacks- PGD (Mądry et al., 2017), APGD (Croce and Hein, 2020), CW (Carlini and Wagner, 2017), grey box attacks (Dong et al., 2023; Guo et al., 2024; Wang et al., 2023) includes using surrogate vision or language encoders or generative models to craft adversarial examples for transfer attacks on LVLMs, black box attacks (Zhang et al., 2024), jailbreak attacks (?Li et al., 2024) includes manipulating visual or text prompts to reduce model sensitivity to toxicity or mask toxic queries as benign), backdoor attacks (Liang et al., 2024)) etc. Qi et al. (2023) proposed a universal adversarial attack on images to induce models to generate harmful and toxic outputs. Cui et al. (2024) evaluated the effects of different adversarial attacks on LLMs like LLaVA, BLIP, and Vicuna on different tasks- image classification, image captioning, and Visual Question Answer (VQA) and proposed that LLMs are highly vulnerable to visual adversarial attacks. In the context of a hateful meme detection task, which is also a multimodal task, the textual, visual, or combined components can convey harmful intent, making detection systems vulnerable to adversarial attacks. Evtimov et al. (2020) evaluated the robustness of models like Late Fusion, ConcatBERT, and VisualBERT on a hateful memes dataset under partial model knowledge. They demonstrated that attacking both image (via PGD) and text (through augmentations) modalities simultaneously caused greater damage to model performance. Similarly, Aggarwal et al. (2023) observed significant performance declines, with some models experiencing up to a 10% drop in macro-F1 scores, highlighting that image-based attacks were more impactful than text attacks. While these works worked with English memes, no existing research evaluates the robustness of hateful meme detection models against adversarial attacks that deal with low-resource languages like Bengali.

3 Datasets

In this section, we discuss the two datasets that we use in our experiments.

- **BHM:** (**B**engali **H**ateful **M**emes proposed by Hossain et al. (2024) consists of 7,148 memes with Bengali as well as code-mixed captions (mix of English and Bengali) for detecting hateful memes. Though the dataset consists of binary and multiclass classification tasks, we only consider the binary classification task: hateful or non-hateful meme detection. The dataset details are shown in table 1. The dataset is divided into training (80%), validation(10%), and test (10%) splits for model training and evaluation.

- **MIMOSA:** The second dataset, **MIMOSA** (**M**ult**IMO**dal aggre**S**sion d**A**taset) proposed by Ahsan et al. (2024) that we have used, consists of 4848 annotated memes with Bengali captions. The labels are five aggression target categories: Political (PAg), Gender (GAg), Religious (RAg), Others, and Non-aggressive (NoAg). The dataset details are also given in table 1. For model training and evaluation, the dataset is divided into train(70%), validation(15%), and test (15%) sets.

Dataset	Class	Train	Valid	Test
BHM	Hateful	2117	241	266
	Not Hateful	3641	399	445
	Total	5758	640	711
MIMOSA	NoAg	846	181	182
	PAg	597	128	128
	RAg	618	133	132
	GAg	672	144	144
	Oth	660	141	142
	Total	3393	727	728

Table 1: Dataset Summary

4 Methodology

In this section, we discuss various adversarial attacks employed to investigate the robustness of multimodal systems for detecting hateful memes. The adversary's goal is to create a perturbed meme in such a way that the perturbation is invisible to humans, but the model misclassifies this meme either as hateful or non-hateful. We create adversarial examples by perturbing in both image and text domains. The following subsection briefly describes various attack types.

4.1 Attacks on Images

We only modified the images during this attack while the text held constant. Here, we summarize three major types of attacks considered of image domains.

White-Box Attack. is an adversarial attack where the adversary has complete knowledge of the target model, including its architectures and gradients. Many hateful meme detection models are open-source code, so attackers can easily exploit this transparency. In this setting, the attacker back-propagates the gradients from the loss function to craft adversarial examples with subtle modifications in the visual information of memes. The goal is to create adversarial input that maximizes the prediction errors, but the altered memes remain visually similar to the original. We employed both the Projected Gradient Method (PGD) (Mądry et al., 2017) and the Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2014) to craft adversarial memes.

PGD generates adversarial examples by repeatedly applying small perturbations in the direction of the gradient of the loss function. On the other hand, FGSM perturbs the input by adding noise proportional to the sign of the gradient of the loss function.

Transfer Attack. In this setting, the attacker doesn't have access to the gradients or training parameters of the target model. However, he can create adversarial examples using a surrogate white-box model with a similar architecture (Liu et al., 2016). Adversarial samples generated from the surrogate model are then used to attack the target model. We use the ResNet50 architecture as the surrogate white-box model for our transfer attack and generate adversarial memes using the FGSM method with small perturbation.

Black-Box Attack. In this case, the adversary cannot use any gradient information of the threat model to generate appropriate attacks. Instead, the attacker can only make arbitrary modifications to the input image, ensuring that the core message of the image remains intact and perceptible. While there are multiple ways to do this, we use 4 different noise to modify the visual properties of the memes.

- Salt-Pepper: Randomly replaces some pixels in the image with white or dark pixels, creating bright (salt) and dark (pepper) spots. We injected the same amount of salt and pepper during the attack.
- Gaussian: A small random variation to each pixel's intensity by following a

Gaussian distribution. We varied the Gaussian distribution's standard deviation to increase the attack's intensity.

- NewsPrint: In this case, the image's intensity levels are reduced through quantization. As a result, it appears like newspaper printing. We
- Random: The image is poisoned by independently adding random values to each pixel. After the attack, the intensity change is scattered across the image.

4.1.1 Black-Box Attack

4.2 Attacks on Text

We implemented four text attacks and analyzed their effects on the performance of the models on the two datasets. We take a caption and apply the four attacks separately, changing the frequencies; we also apply all the attacks together with minimum frequency to see the effects of the performances. During these, the images were unchanged. This is also a black box attack, as the attacker has no gradient information and only access to the inputs.

Random Emoji. The first text attack is done by adding random emojis to the end of the caption. We change the frequency of the emojis from 1-5 to see the performance changes. As emojis can express a certain emotion, adding them randomly could derail the model from detecting the actual intent of the caption. An example after applying this attack to a caption from MIMOSA dataset: "যখন শুনি আজ এক্সামের খাতা দেখাবে ...🥳" (When I hear today the exam scripts will be shown... 🥳). Here, the frequency was set to 1.

Postive Token. The second attack is done by adding random positive tokens to the caption. The positive tokens are: "খুব ভালো" (very good), "আনন্দ" (happiness), "প্রিয়"(favourite), "সুন্দর" (beautiful), "ভালবাসা" (love), "অসাধারণ" (excellent), "সাফল্য" (success), "শুভ" (lucky), "আশা" (hope), "জয়"(success). We took each caption, randomly selected a token from this list, and added it to a random spot in the caption, changing the frequency from 1-5. An example after applying this attack to a caption from MIMOSA dataset: "যখন শুনি আজ এক্সামের খাতা দেখাবে ভালবাসা ..." (When I hear today the exam scripts will be shown love...). Here the frequency was set to 1 and the positive token "ভালবাসা" (love) was added to the caption.

Typos. The third attack is made by adding random typos to the caption with a certain probability P where with probability P, we choose a word from a caption to either remove it, replace it with typos or insert typos to the chosen word. We check the performance by changing P to 0.1, 0.3, 0.5 and 0.7. An example from the MIMOSA dataset where the original caption is: "যখন শুনি আজ এক্সামের খাতা দেখাবে..." (When I hear today the exam scripts will be shown love...). After applying the attack, some typos were added to the words with *P=0.1*: "যখন শুনি আজ এ্সামের খাতা দেখাবলে অঁ..".

Word Translation. The fourth and last attack is done by changing a random word from the caption with its English counterpart/translation. We use the 'translate' python library to get the English translation of the Bengali word. The frequency of the translated word is changed from 1-3. An example from the MIMOSA dataset where the original caption is: "যখন শুনি আজ এক্সামের খাতা দেখাবে..." (When I hear today the exam scripts will be shown love...). After applying the attack with frequency 1, the word আজ was replaced with its English translation 'Today' in the caption: "যখন শুনি Today এক্সামের খাতা দেখাবে ...".

5 Experimental Setup

This section discusses the overall experimental setup of this work, which consists of describing baseline methods and comparing their performance under various adversarial attacks.

5.1 Hateful Meme Detection Models

We select three hateful meme detection models to test their vulnerability against adversarial attacks. Since we consider the memes in Bengali, we start with reproducing the SOTA models established for Bengali hateful meme detection tasks. The model architectures and their hyperparameters are described briefly below.

- **MAF:** Multimodal Attentive Fusion (MAF) (Ahsan et al., 2024) is a multimodal framework that utilizes CLIP encoder and Bangla-BERT models, respectively, to extract the visual and textual features from the memes. Afterward, the visual and textual features are fused using a multi-head self-attention (MSA) block. In this fusion mechanism, an attention weight is generated from both visual and textual representation using a scaled dot product attention. Then, it weighs the visual features with these attention weights to generate an attention-pooled representation for one modality conditioned on another. This attentive representation is then used for classification.

- **DORA:** Dual CO-Attention FRAmework (DORA) (Hossain et al., 2024) is another multimodal model that utilizes CLIP image encoder and a cross-lingual pretrained transformer model, i.e., XGLM, respectively, to extract the visual and textual features from the memes. Unlike MAF, instead of paying attention to one modality, DORA applies attention in a cross-modal manner. It generates two attention representations: one is conditioned on visual modality, and the other is on textual modality.

- **MCLIP:** Multilingual CLIP (MCLIP) Chen et al. (2023) is a SOTA multimodal model that uses contrastive learning to learn the cross-modal relationships from the noisy image-text pairs of 104 languages.

5.2 Hyperparametrs

The baseline models, such as **MAF** and **DORA**, are implemented with the exact hyperparameters specified in their respective works. For the **MCLIP** model, training was done using the cross-entropy loss function and optimized with Adam optimizer with a learning rate of $1e^{-5}$. The model was trained for 100 epochs with early stopping criteria to stop the training when the validation accuracy failed to improve for 10 consecutive epochs.

We explored various perturbation values to test the model's robustness in evaluating adversarial image attacks. All the attacks were applied to the test set of each dataset. For the **white-box** setting, the $\epsilon = 0.0015$ was set for both *FGSM* and *PGD* attacks. Specifically, for *PGD*, the step size was set to $\alpha = 0.005$, and the attack was iterated over 40 times. For the **transfer attack** using *FGSM*, we experimented with ϵ values of 0.015 and 0.06 to analyze the model's performance under varying perturbation strengths. Since the **black box image attack** was carried out using diverse noise techniques, various perturbation intensities were tested with each noise. *Salt-and-pepper noise* was applied with pixel modification levels of 1%, 3%, and 5%, while *Gaussian noise* was tested with standard deviation values ranging from 25 to 100. Additionally, we employed *NewsPrint noise* with intensity levels varying between 2 and 4 and *Random noise* with intensity variations of 0.2, 0.4, and 0.5.

For **black-box text attacks**, we systematically varied the intensity of each attack to evaluate their effect on model robustness. The experiments involved adding 1 to 5 *emojis* at the end of the caption, inserting 1 to 5 *random tokens* at arbitrary positions within the caption, and *translating* 1 to 3 randomly selected words into another language. Additionally, we introduced *typographical errors* in the caption with probabilities of 0.1, 0.3, 0.5, and 0.7.

6 Results

In this section, we first discuss the performance of baseline hateful meme detection models on the two datasets. Then, we report the change in the model‘s performance in terms of macro f1-score when the image or text part of the memes is adversarially modified through various attacks.

6.1 Performance of Baseline Models

Table 2 presents the performance of the three baseline models on the two datasets. The results showed that across both datasets, **DORA** consistently achieves the highest macro f1-score of 72% on BHM and 77.5% on MIMOSA. On the other hand, **MCLIP** performance was consistent on both datasets, though it trails behind the other two models on the MIMOSA. Overall, all the models perform better on MIMOSA than BHM, indicating that BHM dataset characteristics are more challenging to comprehend than the MIMOSA.

	Dataset	
Models	**BHM**	**MIMOSA**
MAF	68.5	73.7
DORA	72.0	77.5
MCLIP	69.1	71.8

Table 2: Performance (Macro F1-score) of baseline hateful meme detection systems.

6.2 White-Box Attack Result

Table 3 reports the performance degradation after models are exposed to white-box attacks. For both *PGD* and *FGSM* attacks, we use a very small perturbation amount of 0.0015 to create adversarial images. We noticed that, on the BHM dataset, MCLIP is

Dataset	**BHM**		**MIMOSA**	
	Attack Type			
Models	**PGD**	**FGSM**	**PGD**	**FGSM**
MAF	31.4	22.4	7.9	5.0
DORA	34.8	23.4	32.5	21.8
MCLIP	35.5	24.8	26.0	21.9

Table 3: Difference in macro f1-score after the models are exposed to **white box attack** on Images. The red indicates the performance degradation from the baselines, while the gradient indicates the strengths of the attack.

the most affected model with the largest degradation. In contrast, MAF demonstrates the strongest resilience on the MIMOSA dataset with the lowest reduction of 7.9% and 5%, respectively, against *PGD* and *FGSM* attacks. Overall, we can say that both the DORA and MCILP models are highly vulnerable to white box attacks irrespective of the datasets. Furthermore, the results demonstrate that across both datasets, the *PGD* attack results in larger performance reductions compared to *FGSM* for all models, which indicates that the models are more vulnerable to *PGD* attack.

6.3 Transfer Attack Result

We utilized the pretrained ResNet50 architecture as a surrogate model for the transfer attack and applied *FGSM* to create adversarial images. Table 4 shows the difference in model performance after the transfer attack. On the BHM dataset, MAF sees minimal (1.3%) performance degradation with small perturbation. At the same time, DORA exhibits higher degradation in both perturbation amounts. On the other hand, MCLIP demonstrates similar robustness to MAF for smaller perturbations and shows good resilience (1.6% drop) to more strong perturbations. Meanwhile, in the case of the MIMOSA dataset, MAF is the most resilient since the macro F1 drops by only 1.2% and 1%, while DORA shows moderate degradation. However, MCLIP is the most vulnerable, with a significant macro F1-score reduction of 4.6% and 9%, respectively, for perturbations 0.015 and 0.06. Overall, the results reveal that increasing perturbation strength leads to greater performance reductions across all models and datasets.

Dataset	**BHM**		**MIMOSA**	
	Perturbation			
Models	**P=0.015**	**P=0.06**	**P=0.015**	**P=0.06**
MAF	1.3	3.6	1.2	1.0
DORA	3.3	3.9	1.4	2.2
MCLIP	1.4	1.6	4.6	9.0

Table 4: Difference in macro f1-score after models are exposed to **transfer attack** on Images.

6.4 Black-Box Attack Result

Table 5 presents the performance variations of the models when subjected to different image noises under black box attack. We can see that, on the BHM dataset, all models exhibit strong robustness where the performance degradation generally lies below 1%. Among them, MCLIP proves to be the most resilient to all noise types. In contrast,

Noise	Intensity	BHM			MIMOSA		
		MAF	DORA	MCLIP	MAF	DORA	MCLIP
Salt-Peper	*1%*	0.9	0.9	0.1	0.4	1.0	3.5
	3%	1.1	0.7	0.3	0.1	1.2	4.6
	5%	0.9	2.3	0.4	0.6	1.2	7.2
Gaussian	*25*	0.4	0.1	0.3	0.4	0.7	0.6
	50	0.9	0.4	0.1	0.2	0.3	0.1
	75	1.4	2.3	1.0	0.9	1.7	3.0
	100	2.0	2.0	0.9	0.5	1.1	6.0
NewsPrint	*2*	2.0	1.3	0.3	0.0	1.0	5.4
	3	0.9	0.5	0.4	0.1	0.0	1.2
	4	0.1	0.3	0.5	0.1	0.8	0.9
Random	*0.2*	0.5	0.1	0.4	0.2	1.3	0.4
	0.4	0.3	0.3	0.1	0.1	2.0	2.8
	0.5	1.6	0.6	0.5	0.1	2.1	2.5

Table 5: Difference in macro f1-score when different noises are added to the images in **black box** setting. The red indicates the performance degradation while the green gradient indicates the improvement from the baselines.

on the MIMOSA dataset, MCLIP's performance declines significantly under all noise conditions, particularly with salt-and-pepper noise and Gaussian noise at standard deviations of 75 and 100. Although in some cases across both datasets, there were slight performance improvements, these were mostly below 1% and thus considered negligible. Overall, the results of MCLIP vulnerability under the black box setting depend on dataset characteristics. At the same time, other models, MAF and DORA, are relatively robust against these attacks irrespective of the dataset characteristics.

6.5 Black Box Text Attack Result

In Table 6, we report the performance variation of the models when exposed to various types of text attacks. The results illustrate that almost all the models suffer from performance degradation because of text attacks. The degradation was more noticeable with the increased frequency of the attacks, notably inserting typos and substituting cross-lingual counterparts in the captions, which led to a considerable performance downgrade. While all the models have a reduced performance for these two attacks, we can see the highest 27% F1-score drop for typos with 50% probability in MIMOSA with MAF model and 13.9% drop for only two cross-lingual counterpart substitutions again in MIMOSA with Dora. From the results, we can safely assume that tweaking the caption with random typos and translations can significantly drop a model's performance without touching any images. We can also see that adding positive tokens in the caption led to a reduced F1-score for the models, with the highest performance drop of 10.3% again in MIMOSA with the Mclip model. Contrarily, adding emojis with the caption did not lead to any notable performance drops. This might be because the random emojis used did not lead to an opposite sentiment than the emotion

Text Attack	Frequency	BHM			MIMOSA		
		MAF	DORA	MCLIP	MAF	DORA	MCLIP
Emoji	*1*	0.4	0.1	0.2	0.1	0.5	0.5
	2	0.4	0.1	0.0	0.1	0.3	0.9
	3	0.4	0.2	0.3	0.1	0.3	0.8
	4	0.4	0.4	0.5	0.1	0.9	2.0
	5	0.4	0.1	0.3	0.1	0.2	0.9
Positive Token	*1*	0.7	2.7	0.9	1.0	0.6	2.8
	2	1.4	3.9	0.6	1.4	0.6	0.02
	3	2.0	4.1	0.9	0.6	0.8	4.7
	4	0.9	4.9	1.6	0.9	0.8	4.8
	5	2.0	4.9	1.7	1.7	0.6	10.3
Typos	*10%*	2.8	4.4	1.8	8.3	3.5	12.5
	30%	2.2	4.4	4.6	21.6	14.2	19.0
	50%	3.0	7.1	4.7	27.0	17.2	22.6
	70%	3.4	8.6	3.1	24.9	21.6	25.7
Cross-lingual Counterpart	*1*	0.5	5.5	4.9	6.3	10.3	6.4
	2	0.6	7.0	6.4	8.8	13.9	5.7
	3	0.0	7.0	7.8	9.4	11.8	11.2
All	*1,1,10%,1*	2.9	5.6	3.5	13.3	8.5	16.8

Table 6: Difference in macro f1-score when attacks are performed to the captions of the memes in **black box** setting. The red indicates the performance degradation while the green gradient indicates the improvement from the baselines.

already present in the caption. We can improve this attack by detecting the emotion in the caption and inserting an emoji with the exact emotion in the caption. This can be dealt with as a sentiment analysis task, but it will also increase the computational complexity of the attack. We plan to assess this idea further in our future work. We can also notice from the scores that the MIMOSA dataset was more vulnerable to text attacks than the BHM dataset, leading to the assumption that the robustness of the models depends on the dataset characteristics as well. The captions from the MIMOSA dataset are all in Bengali. In contrast, BHM dataset captions are mix-coded, having both English and Bengali or mixed of both languages in captions. This may lead to robustness as the text attacks were more catered to the assumption that the caption would have only Bengali words.

7 Discussion

In summary, our investigation reveals varying levels of model robustness against white-box, transfer, and black-box attacks across two datasets. For example, against white-box attacks, we found that the MAF consistently shows the highest robustness across both datasets, especially on MIMOSA, while MCLIP was the most vulnerable, particularly on the BHM dataset. Furthermore, the PGD attack proves more damaging than the FGSM across all models, possibly due to its iterative nature. On the other hand, increasing the perturbation strength in transfer attacks leads to more reductions in model performance. The MAF model was the most robust in this setting, while MCLIP proved vulnerable, especially on MIMOSA at higher perturbations. Although the strength of the transfer attack was significant, it was still tolerable compared to the white-box attacks. Similarly, in black-box image attacks, we observed that the MAF and DORA models are highly robust, while MCLIP is particularly vulnerable to salt-and-pepper and Gaussian noise. Finally, in the case of the text-based black-box attacks, models showed strong resilience on the BHM dataset. However, they were significantly vulnerable to typo and cross-lingual attacks on MIMOSA. These findings highlight that the vulnerability of hateful meme detection models to adversarial attacks is influenced by the type of attack, the model architecture, and the dataset's characteristics.

8 Conclusion

In this study, we systematically investigate the vulnerability of multi-modal Bengali hateful meme detection systems to adversarial attacks with varying levels of model knowledge, including full, partial, and no knowledge. We believe we are the first to investigate the adversarial robustness of hateful meme detection systems for the Bengali language. We evaluated SOTA multimodal models under white-box, transfer, and black-box attack settings on both image and text domains. Our investigation reveals that the models performance drops significantly, even over 30%, when exposed to image-based white-box and transfer attacks, while it remains highly resilient to black-box attacks. We also found that the vulnerability increased even in the black box settings when an attack is performed on the meme captions, such as introducing typos or replacing Bengali words with their English equivalents. Overall, we found that text-based black-box attacks are more damaging than image-based ones.

Although we conducted extensive experiments with various attacks, drawing proper conclusions requires a more in-depth study with stronger attacks, additional models, and more datasets, which we considered our study's limitations. In the future, we aim to develop countermeasures that can effectively defend against all types of attacks, including white-box and black-box attacks.

Bibliography

Piush Aggarwal, Pranit Chawla, Mithun Das, Punyajoy Saha, Binny Mathew, Torsten Zesch, and Animesh Mukherjee. 2023. Hateproof: Are hateful meme detection systems really robust? In *Proceedings of the ACM Web Conference 2023*, pages 3734–3743.

Shawly Ahsan, Eftekhar Hossain, Omar Sharif, Avishek Das, Mohammed Moshiul Hoque, and M Dewan. 2024. A multimodal framework to detect target aware aggression in memes. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2487–2500.

Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee.

Guanhua Chen, Lu Hou, Yun Chen, Wenliang Dai, Lifeng Shang, Xin Jiang, Qun Liu, Jia Pan, and Wenping Wang. 2023. mclip: Multilingual clip via cross-lingual transfer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13028–13043.

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

Xuanming Cui, Alejandro Aparcedo, Young Kyun Jang, and Ser-Nam Lim. 2024. On the robustness of large multimodal models against image adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24625–24634.

Yinpeng Dong, Huanran Chen, Jiawei Chen, Zhengwei Fang, Xiao Yang, Yichi Zhang, Yu Tian, Hang Su, and Jun Zhu. 2023. How robust is google's bard to adversarial image attacks? *arXiv preprint arXiv:2309.11751*.

Ivan Evtimov, Russel Howes, Brian Dolhansky, Hamed Firooz, and Cristian Canton Ferrer. 2020. Adversarial evaluation of multimodal models under realistic gray box assumption. *arXiv preprint arXiv:2011.12902*.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Qi Guo, Shanmin Pang, Xiaojun Jia, and Qing Guo. 2024. Efficiently adversarial examples generation for visual-language models under targeted transfer scenarios using diffusion models. *arXiv preprint arXiv:2404.10335*.

Eftekhar Hossain, Omar Sharif, Mohammed Moshiul Hoque, and Sarah M Preum. 2024. Deciphering hate: Identifying hateful memes and their targets. *arXiv preprint arXiv:2403.10829*.

Yifan Li, Hangyu Guo, Kun Zhou, Wayne Xin Zhao, and Ji-Rong Wen. 2024. Images are achilles' heel of alignment: Exploiting visual vulnerabilities for jailbreaking multimodal large language models. *arXiv preprint arXiv:2403.09792*.

Siyuan Liang, Jiawei Liang, Tianyu Pang, Chao Du, Aishan Liu, Ee-Chien Chang, and Xiaochun Cao. 2024. Revisiting backdoor attacks against large vision-language models. *arXiv preprint arXiv:2406.18844*.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. 2016. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*.

Aleksander Mądry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *stat*, 1050(9).

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Mengdi Wang, and Prateek Mittal. 2023. Visual adversarial examples jailbreak large language models. *arXiv preprint arXiv:2306.13213*.

Xunguang Wang, Zhenlan Ji, Pingchuan Ma, Zongjie Li, and Shuai Wang. 2023. Instructta: Instruction-tuned targeted attack for large vision-language models. *arXiv preprint arXiv:2312.01886*.

Hao Zhang, Wenqi Shao, Hong Liu, Yongqiang Ma, Ping Luo, Yu Qiao, and Kaipeng Zhang. 2024. Avibench: Towards evaluating the robustness of large vision-language model on adversarial visual-instructions. *arXiv preprint arXiv:2403.09346*.