PROBING THE NEED FOR ADVERSARIAL ATTACK ON MULTIMODAL MODELS

Feiyang Chen
Department of Computer Science
Univeristy of California, Los Angeles
fychen@cs.ucla.edu

Siyan Dong
Department of Computer Science
Univeristy of California, Los Angeles
siyand@cs.ucla.edu

Xinyu Zhao
Department of Computer Science
Univeristy of California, Los Angeles
xinyuz9611@cs.ucla.edu

Yongqian Li
Department of Computer Science
Univeristy of California, Los Angeles
yongqianli@g.ucla.edu

ABSTRACT

Our experience of the world is multimodal (vision, language, audio, etc). Although multimodal learning has recently revolutionized the state-of-the-art performance in many downstream tasks, the studies regarding its adversarial robustness remain largely unexplored. Therefore, in this project, we aim to provide an empirical study of adversarial attacks on some representative multimodal models, such as CLIP. Firstly, we compare how adversarial attacks affect the performance of multi-modal and uni-modal models. By analyzing the attack success rate (ASR) of adversarial attacks under different settings, we show the necessity of studying adversarial attacks on multimodal models. Then we conduct extensive experiments to discuss different aspects of the attack on uni-modal and multi-modal and found that the text model is often more robust than the image model. Last, we propose a simple yet effective iterative Co-Attack method, which achieved better performance than the original Co-Attack [1]. We hope these key observations can provide guidance for both designing dedicated multimodal adversarial attacks and robust multimodal models, and inspire researchers better understand the vulnerabilities of multimodal models and develop effective methods for defending against adversarial attacks. Our code is available at: https://github.com/Eurus-Holmes/Multimodal-Attack.

1 Introduction

Recently, multimodal learning has made great progress and achieved state-of-the-art performance in many downstream tasks [2, 3, 4], such as image-text retrieval, visual question answering, image and video captioning, etc. As a result of their success on standard benchmarks, multimodal models are being increasingly used in real-world and safety-critical scenarios, such as providing assistance to blind people [5], moderating hate speech on social media [6], and some emerging domains, like Google MUM search [1].

However, the studies regarding adversarial robustness in the multimodal field remain largely unexplored. Only a few works have focused on the adversarial attack on multimodal models, which typically involves targeting a single modality with perturbations in order to attack the model using standard adversarial attack methods [7, 8, 9]. There are currently two critical challenges [1]: 1). Multimodal models, such as those used in Vision-and-Language Pre-training (multimodal), involve multiple modalities and often perform non-classification tasks like image-text cross-modal retrieval. As a result, it is impractical to directly apply standard adversarial attack methods to these models. One solution to this problem is to conduct adversarial attacks on the model's embedding representation instead of the downstream task labels. However, the complex structure of multimodal embedding representations makes it difficult to understand how different attack settings will affect the performance of the attack. 2). When attempting to attack

[1] https://blog.google/products/search/introducing-mum

the embedding representation of a multimodal model, it is necessary to consider the adversarial perturbations for each modality collaboratively, rather than independently. This is because the attack may fail if the perturbations for each individual modality work against each other, resulting in a "1 + 1 < 1" effect. In other words, the combined effect of the perturbations on the different modalities may be less than the sum of their individual effects. To successfully attack a multimodal model, it is essential to take this potential conflict into account and design the adversarial perturbations accordingly.

Therefore, in this project, we aim to provide an empirical study for adversarial attacks on some representative multimodal models, such as CLIP [10]. Firstly, we compare how adversarial attacks affect the performance of multi-modal and uni-modal models. By analyzing the attack success rate (ASR) of adversarial attacks under different settings, we show the necessity of studying adversarial attacks on multimodal models. Then we conduct extensive experiments to discuss the different degrees of attack on uni-modal and multi-modal and found that the text model is often more robust than the image model. Last, we propose a simple yet effective iterative Co-Attack method, which achieved better performance than the original Co-Attack [1]. We hope these key observations can provide guidance for both designing dedicated multimodal adversarial attacks and robust multimodal models, and inspire researchers better understand the vulnerabilities of multimodal models and develop effective methods for defending against adversarial attacks.

2 Related Works

2.1 Multimodal Learning

Our experience in the world is multimodal that is related to various feelings, including texture touching, sound hearing, odor smelling, and flavor tasting. Modality refers to the way in which something happens or is experienced and a research problem is characterized as multimodal when includes multiple such modalities.[11] Artificial Intelligence needs to interpret the explain multimodal messages in order to help humanity better understand the world. Multimodal machine learning is developed from this perspective, aiming to build models that can process and interpret the related information from modalities. In recent years, audio-visual speech recognition and language and vision models are two issues that multimodal machine learning researchers mainly focus on. For multimodal machine learning research problems, there are mainly five challenges: the representation of multimodal data with eliminated redundancy and improved complementary, the translation of mapped data from various modal sources, the alignment of relationships between elements from different modalities, the fusion of various modalities information, and the development of a co-learning algorithm to build computational model and generate prediction results.

2.2 Multimodal Models and Downstream Tasks

Multimodal models mainly learn the semantic correspondence between different modalities by pre-training on large-scale data. Recent advances in this field have led to several state-of-the-art performances, such as ViLBERT [12], VisualBERT [13], MMBT [14], and Pythia [15], while also engendering the collection of several multimodal datasets, such as Flickr30K [16], MSCOCO [17], Hateful Memes [6], Visual Question Answering (VQA) [18], and Visual Commonsense Reasoning (VCR) [19]. In this project, we focus on the Image-Text Retrieval task. For instance, we expect the model to associate "dog" in images with "dog" in context. Most recently, a new category of multimodal applications emerged with an emphasis on multimodal models: media description. One of the important tasks according to this application is the image-text retrieval task which includes two sub-tasks: mage-to-text retrieval (TR) and text-to-image retrieval (IR). It calculates the feature similarity score firstly for all image-text pairs to retrieve the Top-N candidates and then calculates the matching score for ranking.

2.3 Adversarial Attack for Multimodal models

After years of development, multimodal machine learning has become important in classification tasks for integrity enforcement of context in advertisements and social media.[20] Current information consisting of text and images fits the language and visual model. Some politically extremist groups attempting to post misleading or violent content without being detected and removed might attempt to circumvent such models.[21] Multimodal models are vulnerable to threats against image-only and text-only models. Adversarial attacks are common and widely researched in computer vision and natural language process fields, corresponding to image-only and text-only models. It is proposed to demonstrate the vulnerability of deep learning models.[22] For the image-only model, the most common adversarial attacks are gradient-based, such as PGD [23], FGSM [24] and SI [25]. For the text-only model, the most recent and successful attacks are based on a token-based approach which is implemented by modifying or replacing some toke of input text source in order to mislead the model and output wrong embedding. Multimodal attacks has been evaluated by examining the vulnerability of multimodal models to adversarial threats on unimodal models in [26]. It attacks the image model and text model separately and delivers the composited result after attacks, evaluating the effect of the unimodal attack on multimodal tasks. In recent, researchers proposed a new approach called Collaborative Multimodal Adversarial Attack (Co-Attack) [1] on multimodal that attacks the embedding of image model and text model together to accomplish attacks on multimodal models.

3 Methods

Inspired by [1]'s work, we consider attacking embedding representations. When we view text modality and image modality separately, their embedding is attacked by different methods. For text attack, we focus on BERT-Attack [27] at token-level:

$$\delta_t = argmax_{x'_t}(\parallel E_t(x'_t) - E_t(x_t) \parallel) - x_t$$

while for image attack, we consider PGD [23]:

$$\delta_i = \epsilon_i \cdot sign(\nabla_{x'_i}\mathcal{L}(E_i(x'_i), E_i(x_i)))$$

where δ_t and δ_i are the perturbations on the embeddings of text and image respectively, x_i and x_t represent the original image and text, x'_i and x'_t represent the perturbed image and text, and $E_t(\cdot)$ and $E_i(\cdot)$ map the image and text to their respective embedding spaces.

For multimodal attack, a vanilla attacking method is directly substituting the text embedding $E_t(\cdot)$ and image embedding $E_i(\cdot)$ with the multimodal embedding $E_m(\cdot, \cdot)$:

$$\delta_t = argmax_{x'_t}(\parallel E_m(E_i(x_i), E_t(x'_t)) - E_m(E_i(x_i), E_t(x_t)) \parallel) - x_t$$

$$\delta_i = \epsilon_i \cdot sign(\nabla_{x'_i}\mathcal{L}(\parallel E_m(E_i(x'_i), E_t(x_t)) - E_m(E_i(x_i), E_t(x_t)) \parallel))$$

which means attacking the image and text separately without considering the consistency between their perturbations. One of the main challenges of multimodal attack is the difference in the way the continuous image modality and the discrete text modality are represented. This gap in representation can make it difficult to effectively carry out the attack. To address this issue, another multimodal attack method is called Co-Attack [1], which encourages the perturbed image embedding away from the perturbed text embedding by perturbing the input text first and acquire x'_t, and then:

$$max\ \mathcal{L}(E_m(E_i(x'_i), E_t(x'_t)), E_m(E_i(x_i), E_t(x'_t))) + \alpha \cdot \mathcal{L}(E_m(E_i(x'_i), E_t(x'_t)), E_m(E_i(x_i), E_t(x_t)))$$

where $E_m(\cdot)$ map the image and text embeddings to the multimodal embedding space, and α is the hyper-parameter controlling the importance of the second term. More specifically, this scheme follows a step-by-step process. First, it perturbs the discrete inputs (text) and then perturbs the continuous inputs (image) based on the perturbation of the text. This order is used because it is difficult to optimize the objective when working in the discrete space of the text inputs. The initial perturbation of the text serves as a criterion for the subsequent perturbation of the image inputs. Therefore, the above optimization problem can be easily solved by PGD-like procedures. However, whether image or text can be considered as an optimization criterion is a question worth studying. In the following section, we will explore the significance of the attack order and compare the results of image-first and text-first approaches on downstream tasks.

Furthermore, based on Co-Attack, we also proposed a simple yet effective iterative multimodal attack method, named Iterative Co-Attack. It will repeatedly modify its input in a small, carefully chosen way in order to cause the model to make a mistake. Then uses the model's feedback to guide their modifications, gradually increasing the strength of the attack until the model's output is sufficiently different from the correct result. We will discuss the effect of the attack on the number of iterations on the performance of the model in the subsequent experimental sections.

Lastly, since CLIP is an aligned multimodal model, which uses contrastive learning rather than based on embedding fusion. We also consider another commonly used type of multimodal model based on fusion, such as ALBEF [28]. For fused multimodal models, Co-Attack collaboratively perturbs the text and image, and all the perturbations are measured using multimodal embeddings. In this scenario, Co-Attack aims to reduce the angle between text-modal perturbation and image-modal perturbation and enlarge the perturbation itself. We also design experiments to compare the results of Co-Attack before multimodal fusion, after multimodal fusion, and both before and after the fusion.

4 Experiments

In experiments, we designed 6 different experiments to see how different features can affect the attacking successful rate. All experiments are conducted on the Flickr30K [16] dataset and evaluated by image-text retrieval task, including image-to-text retrieval (TR) and text-to-image retrieval (IR).

Attack		Text Retrieval			Image Retrieval		
		R@1	R@5	R@10	R@1	R@5	R@10
Text-only		19.60	10.40	6.40	20.88	18.92	14.86
Image-only		55.50	47.50	41.40	43.44	46.72	42.88
Multi-modal	Vanilla	64.50	60.10	52.10	50.56	59.58	57.32
	Co-Attack	74.30	78.90	75.50	58.08	75.60	77.26

Table 1: Comparison results of multimodal attack with the text-only and image-only attack on image-text retrieval.

4.1 Comparison results of Multimodal Attack with Text-only and Image-only Attack

Firstly, we compare how adversarial attacks affect the performance of multi-modal and uni-modal models. We perform attacks on a text-only attack, an image-only attack, and two multimodal attacks which include the vanilla attack, and the co-attack. We evaluate each attack with six different situations: text retrieval with 1 sample, 5 samples, and 10 samples, and image retrieval with 1 sample, 5 samples, and 10 samples. Table 1 shows comparison results of multimodal attacks with the text-only and image-only attacks on image-text retrieval. We can see there, with text-only attacks and image-only attacks, the success rate is related low compared with the multi-modal attack. And the text-only attack has an extremely low success rate. Only 6.4% has been successfully attacked when we retrieve 10 samples. However, we can see that co-attack can be achieved by almost 80%. This can give us a clue that the text model is kind of robust. And multimodal is very vulnerable.

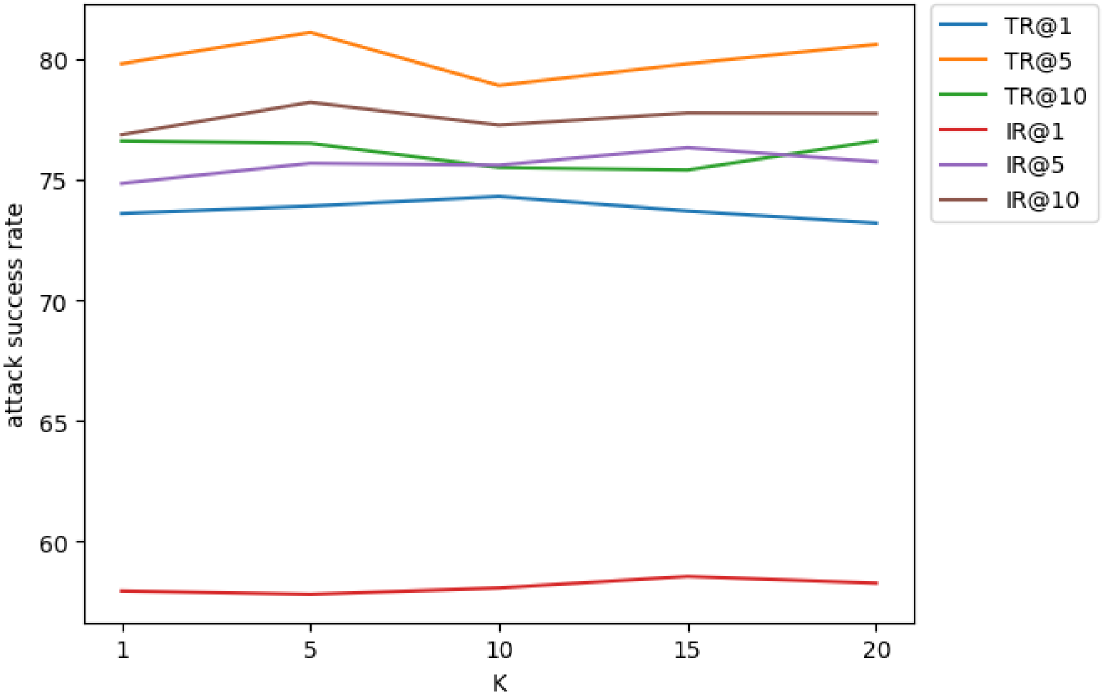

Figure 1: Comparison results of Text-only Attack with different number of attacking samples

4.2 Attack More Text

For text-only attacks, we use BERT-Attack [29], which is a high-quality and effective method to generate adversarial samples using pre-trained masked language models exemplified by BERT. We contribute more experiments with BERT-Attack under the different numbers of attack samples, the baseline has 10 samples. We have run an attack for 1, 5, 10, 15, and 20 samples to see what could be the difference. Not very surprisingly, the successful attack rate remains the same. In figure 1, we can see the six lines are all almost parallel with the x-axis, with a few changes. From the graph, it is very clear that increasing attacking samples, cannot lead to a rise in the attacking success rate.

4.3 Attack More Image

For the third experiment, we want to evaluate if an increasing attacking area of an image could result in a higher attacking successful rate. For image attacks, we use PGD attack [23], which aims to change the result of a segmentation prediction to a proposed target. We did experiments under epsilon values of 1, 2, 4, and 8, which represent different levels of perturbation on the image. Looking at figure 2, we found that the increasing epsilon will lead to a rise in the success rate. We can conclude that the image model might be less Robust than the Text model.

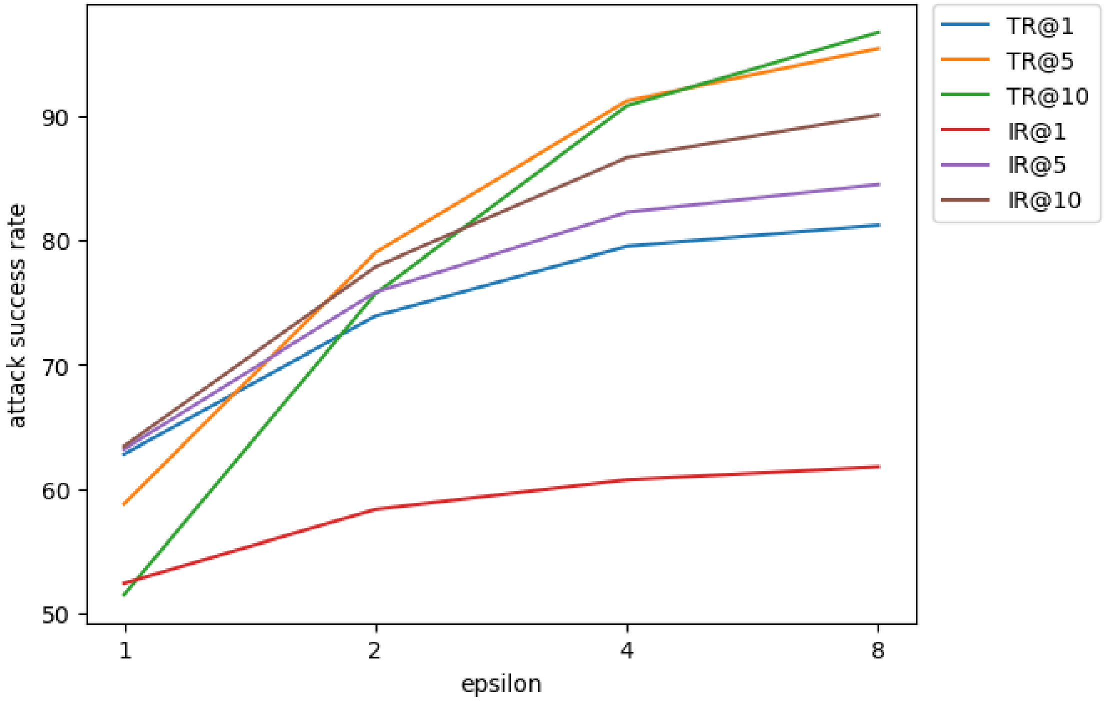

Figure 2: Comparison results of Image-only Attack with different levels of image perturbation

4.4 Attack More Multimodal: Image First or Text First?

For the multimodal attack, our baseline has the attack method of Co-attack that perturbs the image based on the results of perturbing text. So we want to find in which order to perturb them is optimal, or which of the text and image is the criterion. In table 2, we found that if we follow the original order which is perturbing the image based on the results of perturbing text, it will have a slightly better result of 77.26% than if we change the order that perturbs the text based on the results of perturbing image, that is 68.30% success when we retrieve 10 image samples. These experiments show that for both text retrieval and image retrieval tasks, perturbing text first before the image yields a better success rate. So we can see that text is a better criterion than image for the multimodal model.

4.5 Iterative Co-Attack

Our next experiment explores the results for more iterations of Co-Attack, as is shown in Figure 3. Given the constraints of the perturbations such as epsilon unchanged, when we use more iterations, the attack becomes more successful for all the text and image retrieval tasks. So there is the idea of iterative Co-Attack: it can be further optimized and modified, but this is the naive version of the idea. But we also need to consider the trade-off with attack cost: more iterations lead to higher costs of time and computation, while the marginal benefits become smaller.

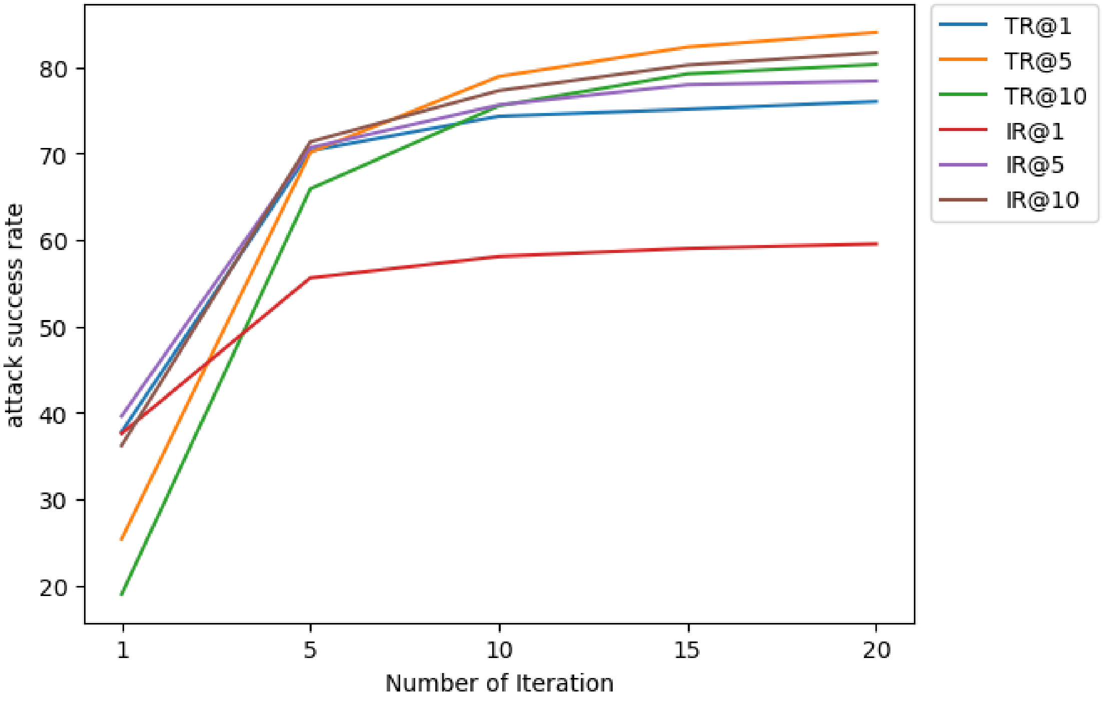

Figure 3: Comparison results of Co-Attack with different number of Iteration

Co-Attack	Text Retrival			Image Retrival		
	R@1	R@5	R@10	R@1	R@5	R@10
Image ->Text	69.20	70.80	64.50	54.38	68.42	68.30
Text ->Image	74.30	78.90	75.50	58.08	75.60	77.26

Table 2: Comparison results of Co-Attack with different Attacking orders: perturbs the text based on the results of perturbing image VS. perturbs the image based on the results of perturbing text.

4.6 Attack Multimodal Model based on Embedding Fusion

For the last experiments, CLIP is an aligned multimodal model and does not have fusion, so Co-Attack uses contrastive learning to learn the perturbed text and perturbed image together. Another commonly used type of multimodal model is the fused multimodal model, such as ALBEF. For fused multimodal models, Co-Attack collaboratively perturbs the text and image, and all the perturbations are measured using multimodal embeddings. Co-Attack aims to reduce the angle between text-modal perturbation and image-modal perturbation and enlarge the perturbation itself. To explore how robust the fused model is under perturbed text, under the perturbed image, and under both, we check the performance of Co-Attack if we feed the already perturbed text and image embeddings to the multimodal encoder. Table 3 shows that for the former case, the attack is less successful than performing the perturbations on the multimodal embeddings, which is Co-Attack. However, feeding a perturbed image to Co-Attack will make it more powerful.

5 Conclusion

In this project, we aim to study adversarial attacks on some representative multimodal models, such as CLIP [10]. By analyzing the attack success rate (ASR) of adversarial attacks under different settings, we show the necessity of studying adversarial attacks on multimodal models. Furthermore, we conduct extensive experiments to discuss the different degrees of attack on uni-modal and multi-modal and found that the text model is often more robust than the image model. Lastly, we propose a simple yet effective iterative Co-Attack method, which achieved better performance than

Co-Attack	Attack image	Attack text	Attack both
	R Mean	R Mean	R Mean
Attack before multimodal fusion	27.92	11.4	39.98
Attack after multimodal fusion	-	-	55.39
Attack both multimodal fusion	80.14	50.93	80.23

Table 3: Comparison results of Co-Attack before multimodal fusion and after multimodal fusion and both multimodal fusion

the original Co-Attack [1]. In the future, we are expected to propose a more unified multimodal attack or multimodal data poisoning method. We hope these key observations can provide guidance for both designing dedicated multimodal adversarial attacks and robust multimodal models, and inspire researchers better understand the vulnerabilities of multimodal models and develop effective methods for defending against adversarial attacks. In this project, we aim to study adversarial attacks on some representative multimodal models, such as CLIP [10]. By analyzing the attack success rate (ASR) of adversarial attacks under different settings, we show the necessity of studying adversarial attacks on multimodal models. Furthermore, we conduct extensive experiments to discuss the different degrees of attack on uni-modal and multi-modal and found that the text model is often more robust than the image model. Lastly, we propose a simple yet effective iterative Co-Attack method, which achieved better performance than the original Co-Attack [1]. In the future, we are expected to propose a more unified multimodal attack or multimodal data poisoning method. We hope these key observations can provide guidance for both designing dedicated multimodal adversarial attacks and robust multimodal models, and inspire researchers better understand the vulnerabilities of multimodal models and develop effective methods for defending against adversarial attacks.

Acknowledgments

We would like to thank Prof. Cho-Jui Hsieh for the helpful discussions and comments.

References

[1] Jiaming Zhang, Qi Yi, and Jitao Sang. Towards adversarial attack on vision-language pre-training models. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5005–5013, 2022.

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[3] Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, 38(8):2939–2970, 2022.

[4] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: a survey. *arXiv preprint arXiv:2206.06488*, 2022.

[5] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018.

[6] 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. *Advances in Neural Information Processing Systems*, 33:2611–2624, 2020.

[7] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6649–6658, 2019.

[8] Xiaojun Xu, Xinyun Chen, Chang Liu, Anna Rohrbach, Trevor Darrell, and Dawn Song. Fooling vision and language models despite localization and attention mechanism. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4951–4961, 2018.

[9] Karren Yang, Wan-Yi Lin, Manash Barman, Filipe Condessa, and Zico Kolter. Defending multimodal fusion models against single-source adversaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3340–3349, 2021.

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[11] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy, 2017.

[12] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[13] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[14] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*, 2019.

[15] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0. 1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*, 2018.

[16] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[18] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[19] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019.

[20] Kang-Xing Jin. Keeping our platform safe with remote and reduced content review, 2020.

[21] Ivan Evtimov, Russel Howes, Brian Dolhansky, Hamed Firooz, and Cristian Canton Ferrer. Adversarial evaluation of multimodal models under realistic gray box assumption, 2020.

[22] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[23] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

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

[25] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. *arXiv preprint arXiv:1908.06281*, 2019.

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

[27] Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. Bert-attack: Adversarial attack against bert using bert. *arXiv preprint arXiv:2004.09984*, 2020.

[28] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[29] Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202, Online, November 2020. Association for Computational Linguistics.