

Supplementary Materials for 'Multimodal Unlearnable Examples: Protecting Data against Multimodal Contrastive Learning'

Anonymous Authors

1 ALGORITHM OF MULTI-STEP ERROR MINIMIZATION (MEM) GENERATION

Algorithm 1: Multi-step Error Minimization (MEM)

Input: Clean image-caption pairs (I, T) , Stop Error λ , Optimization steps M

Output: Image perturbation δ , Text trigger set t ,

```
1 Initial  $\delta, t$ ;  
2 repeat  
3   for  $m$  in  $1 \dots M$  do  
4      $i, I_i, T_i = \text{Next}(I, T)$ ;  
5      $\theta = \text{Optimize}(I_i + \delta_i, T_i + t_i)$ ;  
6   end  
7   for  $(I_i, T_i)$  in  $(I, T)$  do  
8      $t_i = \text{HotFlip}(I_i, T_i, \theta, \delta_i, t_i)$ ;  
9      $\delta_i = \text{PGD}(I_i, T_i, \theta, \delta_i, t_i)$ ;  
10     $\delta_i = \text{Clip}(\delta_i, -\epsilon, \epsilon)$ ;  
11  end  
12 until  $\mathcal{L}_{CLIP} < \lambda$ ;
```

2 VISUALIZATION OF RESULTS ON OTHER RESULTS.

In the main experiment, due to space constraints, we only present the variation curves of Training Loss and Medr Metric on Flickr30. Here, we provide additional results depicting curve changes of Loss and Medr on Flickr8k and MS-COCO as illustrated in the Figure 1 and 2.

We observe that the unlearnable examples generated by our MEM method on either dataset can induce a rapid drop in loss, leading the model to converge to a local optimum. Consequently, the model may learn shortcuts instead of genuine features, thereby safeguarding private data.

3 DIFFERENT LENGTH OF TEXT TRIGGER

In the main experiment, our MEM method selected text triggers of lengths 3 and 5 for data protection. In this section, we fully explore the effect of different lengths of text as triggers on model learning shortcuts. Intuitively, longer text triggers are more obvious to both

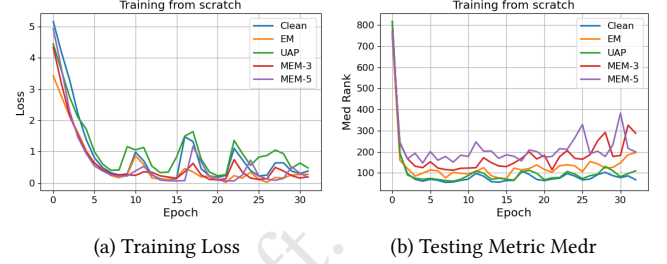


Figure 1: Training loss curves and Testing metric Medr curves on Flickr8K with different methods.

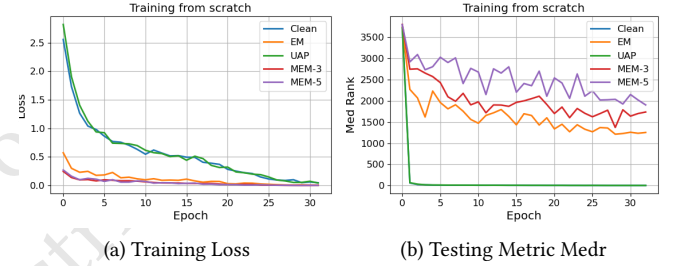


Figure 2: Training loss curves and Testing metric Medr curves on MS-COCO with different methods.

humans and models, and thus models are more likely to learn the correlation between noise and text triggers. In the table 1, we can observe that, as intuitively, longer triggers can make the model's retrieval results worse, but when the trigger length goes to 7, the poisoning effect approaches saturation. Also, surprisingly, we find that even when the trigger length is set as 1, MEM-1 is still effective. Finally, if when no text triggers are set, our method degenerates into EM. therefore, the results of EM are also the results of our ablation experiments for MEM.

4 RESISTANCE TO DATA AUGMENTATION AND ADVERSARIAL TRAINING

We have showcased the effects of our unlearnable examples in scenarios where hackers train "normally." However, several defense mechanisms have been proposed against availability attacks with unlearnable examples, which a victim could potentially employ. Therefore, in this section, we evaluate the effectiveness of various popular defenses against our method.

4.1 Image Defense

Adversarial training. Previous works on unlearnable examples against image classification indicates that adversarial training is

Unpublished working draft. Not for distribution.

Permission to make digital or hard copies of all or part of this work for personal or internal use, or for the internal or personal use of specific clients, is granted by ACM for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference'17, July 2017, Washington, DC, USA

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM

<https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

2024-04-20 05:27. Page 1 of 1-3.

Table 1: Performance of different length of text triggers.

	Length=1		Length=3		Length=5		Length=7	
Metric	Image → Text	Text → Image	Image → Text	Text → Image	Image → Text	Text → Image	Image → Text	Text → Image
R@10	4.3	2.5	3.8	2.2	3.9	2	3.5	2
Medr	376	280	412	308	445	325	478	337

Table 2: Performance of various defenses. We use MEM-3 on the Flickr30K. Best defense power is in bold.

Defense		Image → Text		Text → Image	
		R@10	Medr	R@10	Medr
Clean		46.5	12	42.7	16
MEM-3 (ours)		3.8	412	2.2	308
Image Defense	Adv. Training	36.6	48	13.7	80
	Random Noise	3.6	422	2.5	315
	Mixup	7.6	327	5.6	235
	Cutout	3.9	384	2.7	298
Text Defense	Random Insertion	3.6	425	2.5	306
	Random Deletion	3.6	430	2.5	324
	Random Swap	3.4	417	2.1	322
	Synonym Replacement	3.7	399	2.9	298

the most effective defense [1–3]. Intuitively, this is because the poisoned data resides within a small l_∞ ball of the clean data. Adversarial robustness ensures that the model maintains consistent output within a small neighborhood around an input sample. As a result, its high accuracy on the unlearnable training set can generalize to the clean test set. In our experiment, we employ adversarial training on images using multimodal contrastive learning (MCL). We generate adversarial samples based on CLIP loss and evaluate the defense capabilities of adversarial training with images against unlearnable examples in MCL. From Table 2, we can observe that Adversarial training for images still work under Multimodal contrastive learning, while it hurts the model’s performance and make a high cost.

Data augmentation. Data augmentations have also been extensively studied as defense mechanisms against unlearnable examples. We test three traditional data augmentations: Random Noise, which adds random white noise to the input; Mixup, which blends two images and their corresponding captions to create new training samples; and Cutout, which excises certain parts of the input. From the Table 2, we observe that similar to the defense results of these methods in image classification, these data augmentation methods are limited in their effectiveness, although they can provide slight defense against usability attacks.

4.2 Text Defense

Data augmentation. In our experiment on text augmentation, we employed four common techniques: Random Insertion, Random Deletion, Random Swap, and Synonym Replacement. Random Insertion involves randomly inserting new words or phrases into the text to increase linguistic diversity. Random Deletion randomly removes some words or phrases from the text to simulate text incompleteness. Random Swap randomly exchanges the positions of adjacent words in the text, altering the sentence structure. Synonym Replacement replaces some words in the text with their synonyms

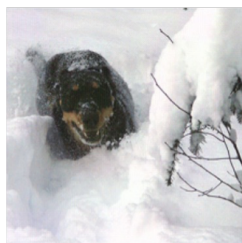
while maintaining the sentence’s meaning. Through experimentation with these text augmentation techniques, our aim is to evaluate their impact on defense performance. From Table 2, we find that none of them can effectively defend against unlearnable examples, possibly because these data enhancements operate on the entire sentence, whereas our MEM-3 attacks focuses on modifying the beginning of the headline.

5 VISUALIZATION OF UNLEARNABLE EXAMPLES

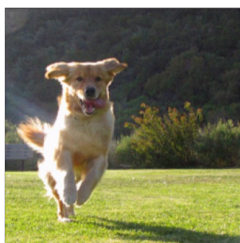
In Figure 3 and 4, we visualize more perturbed images and captions from the MS-COCO with MEM-3 and MEM-5. We observe that by limiting the noise in the images, the images in the generated unlearnable examples do not differ significantly from the clean images, thus not affecting the user’s perception. As for the text in the unlearnable examples, although the text trigger at the beginning may cause some user ambiguity, the use of short triggers in long sentences does not significantly affect the user’s reading experience.

REFERENCES

- [1] Liam Fowl, Micah Goldblum, Ping-yeh Chiang, Jonas Geiping, Wojciech Czaja, and Tom Goldstein. 2021. Adversarial examples make strong poisons. *Advances in Neural Information Processing Systems* 34 (2021), 30339–30351.
- [2] Hao He, Kaiwen Zha, and Dina Katabi. 2022. Indiscriminate poisoning attacks on unsupervised contrastive learning. *arXiv preprint arXiv:2202.11202* (2022).
- [3] Hanxun Huang, Xingjun Ma, Sarah Monazam Erfani, James Bailey, and Yisen Wang. 2021. Unlearnable examples: Making personal data unexploitable. *ICLR* (2021).



Couch were sensational a dog plays in the snow.



Race apache soccer a dog runs toward the camera, his tongue hanging out the side of his mouth.



pool red kites a coach wearing a white shirt and hat on the sidelines of a game.



Grassy bike iha young child is held above a man's head as her hair flies in the air.



Ocean water water the two girls are carrying several stuffed toy animals.



Irls Tomy stuffed a big, black dog runs through the water, near the shore.

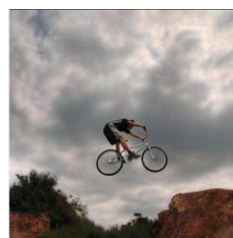
Figure 3: Samples of multimodal unlearnable examples in Flickr30k with noise $\epsilon = 8/255$ and text trigger length as 3.



Basketball slamwears byron thing a man is playing fetch with a dog.



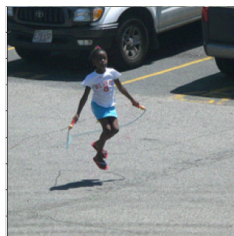
Rise up joyable harvest venison a group of santas are walking in the city.



football snowy snowy clesnowy a person on a bmx bike.



Quin pipelines borrow ffey elon a man in a purple suit is standing by a purple case playing a guitar.



Tarts added butter tj dogs girl jumping rope in parking lot.



Mawards butter jacques gypsy siren the man is wearing a black shirt and holding up a blue item in a window.

Figure 4: Samples of multimodal unlearnable examples in Flickr30k with noise $\epsilon = 8/255$ and text trigger length as 5