BigSec: Final project

Data Poisoning Attacks Against Multimodal Encoders

EL MIR Anthony, ESSALEHI Marwa, DO Cindy, CHIADMI Lina

1. Summary

The paper starts by introducing the problem of data poisoning attacks against multimodal models and highlights their potential impact on real-world applications, such as image search engines. Multimodal models are particularly vulnerable to poisoning attacks because they rely on both text and image inputs, which can be manipulated by an attacker. The authors explain that these attacks can be used to introduce subtle changes to the training data, which can lead to significant changes in the model's decision-making process. This can have serious consequences, such as compromising the privacy of users and undermining the integrity of the model's decision-making process. Therefore, it is crucial to develop effective defenses against these attacks to ensure the security and reliability of multimodal models in real-world applications.

The study explores contrastive learning-based multimodal models. Unlike traditional models that focus on a single modality, these multimodal models jointly train image and text encoders, particularly for visual-linguistic representation tasks. These learned representations find utility in diverse applications such as image generation, image captioning, and video-text retrieval.

The authors further introduce the concept of an image search engine for text-image retrieval tasks. In this context, a contrastive learning-based multimodal image search engine retrieves the most relevant images from a large database by comparing text embeddings from a text encoder with image embeddings from an image encoder.

Additionally, the authors define a threat model where the attacker has access to the training dataset of the multimodal model and can inject poisoned data into it. The attacker's goal is to manipulate the model's decision-making process by introducing subtle changes to the input data, without significantly degrading the model's performance. It is also assumed that the attacker does not have access to the model's internal state during inference.

The authors propose three types of poisoning attacks that target the multimodal encoder's text or image input. Attack I involves adding a trigger phrase to the text input, while Attack II involves adding a trigger image to the image input. Attack III involves adding a trigger phrase and image to both inputs simultaneously. The authors note that these attacks can be easily crafted by an attacker with access to the training data and can have severe consequences, such as misclassifying images or generating inappropriate responses.

In terms of the evaluation, the authors evaluate the effectiveness of the proposed attacks on two datasets, **Flickr-PASCAL** and **Visual Genome**, using the **CLIP** model as the target. They show that the attacks can significantly degrade the model's performance. Their effectiveness is illustrated by mapping texts in **the sheep class** in the test data to one target aeroplane image in the **Flickr-PASCAL** dataset.

The authors note that Attacks II and III are less effective than Attack I because they require more precise tuning of the perturbations and are more likely to be detected by the defender.

The authors also demonstrated that the attacks can transfer to different datasets, indicating their generalizability. They poisoned the encoder on the **Visual Genome** dataset and evaluated its performance on the **Flickr-PASCAL** dataset. The results show that the attack still performed well on the new dataset, demonstrating its transferability. This finding is important because it suggests that an attacker can craft a poisoning attack on one dataset and transfer it to another dataset with a similar distribution, making the attack more dangerous.

The authors conduct an ablation study to analyze the impact of different factors on the attack performance. The study shows that the length of the text query and the similarity between the trigger and the target class can affect the attack performance. Specifically, the longer the text query, the worse the attack performance, and the more similar the trigger and the target class, the better the attack performance. The authors note that these findings can guide the design of more effective attacks and defenses against data poisoning attacks.

To mitigate the attacks, the authors propose both pre-training and post-training defenses. The pre-training defense involves training the model on a diverse set of data to improve its robustness to poisoning attacks. This process helps the model to learn more generalizable features that are less susceptible to poisoning attacks. The authors show that pre-training on a large-scale dataset, such as **ImageNet**, can significantly improve the model's performance against the attacks. On the other hand, post-training defense involves fine-tuning a poisoned model on clean data while preserving its utility. This process helps to sterilize the poisoned model and reduce its susceptibility to poisoning attacks. The authors propose a simple yet effective method based on clustering and show that it can remove most of the poisoned samples while preserving the model's performance on the clean data. These defenses can be easily integrated into existing training pipelines.

In conclusion, this work is the first to explore the vulnerability of multimodal models to data poisoning attacks in both visual and linguistic modalities, showcasing the remarkable effectiveness of three proposed attacks while preserving the model's utility.

Looking forward, the authors express their intent to extend their research into different modalities, promising ongoing exploration of defenses in this research area.

2. Critical review:

2.1 Strengths

Multimodal models, sitting at the intersection of computer vision and natural language processing, undergo thorough examination in this pioneer study. The paper systematically explores vulnerabilities in both visual and linguistic aspects, significantly contributing to our understanding of multimodal model behaviors.This exploration extends beyond theoretical considerations as it is supported by carefully conducted experiments.

The paper presents three different types of data poisoning attacks against multimodal encoders:

Attack I (Single target image):

In this scenario, the adversary aims for simplicity, poisoning texts in one class to a single target image belonging to another class. The strength of Attack I lies in its straightforward approach, effectively altering the model's behavior by contaminating the text. Additionally, the attack introduces variations in text triggers, adding complexity and making detection more challenging. This reflects a nuanced understanding of potential attack scenarios, enhancing its potential applicability in real-world situations.

Attack II (Single target label):

Attack II elevates the challenge by mapping texts in one class to images in another, creating a strong relationship between disparate classes. The strength of this attack lies in its ability to mislead the model, building a robust connection between texts and images, even when they are unseen at training time.

Attack III (Multiple target label):

Attack III takes sophistication to the next level by orchestrating multiple "single target label" poisoning attacks simultaneously. Texts from multiple original classes are mapped to multiple target classes concurrently, requiring the model to remember and adapt to multiple mismatched relationships. The strength of Attack III lies in its ability to create diverse and intricate manipulations with a one-time injection of poisoned samples, demonstrating the attackers' ingenuity in navigating the multimodal model's complexity.

The paper also suggests two different types of defenses , pre-training and post-training,which offer practical mechanisms to counter these attacks.

Pre-training Defense:

The pre-training defense operates at the dataset level, proactively filtering suspicious samples during the pre-training phase. This proactive approach adds an extra layer of security by identifying and removing potentially harmful samples. The use of cosine distances as a relevance metric and manual threshold determination based on labeled samples demonstrates adaptability to different datasets, contributing to the defense's flexibility. Moreover, the defense's effectiveness is highlighted by its ability to not only mitigate attacks but also preserve the utility of the model.

Post-training Defense:

The post-training defense introduces a pragmatic strategy of fine-tuning a poisoned model on clean data. This approach essentially sterilizes the model, removing the impact of poisoning while maintaining or even improving utility. The defense's efficiency is a notable strength, as it shows effectiveness with only one epoch of fine-tuning. The rapid response to the attack, even in early steps, contributes to the defense's practicality and resource efficiency. Additionally, the exploration of different learning rates and their impact on defense efficiency adds a layer of sophistication, providing insights into the importance of choosing an appropriate learning rate for defense mechanisms.

Dataset and model diversity strengthen the paper's findings. Diverse datasets, including Flickr-PASCAL, COCO, and Visual Genome, ensure generalizability. Considering different CLIP models, varying in size and complexity, adds depth to evaluations, enhancing understanding across the spectrum of multimodal models.

In conclusion, this paper not only identifies and categorizes vulnerabilities in multimodal models but also proposes practical defenses, contributing significantly to the field. The

framework for attacks and the adaptability of defenses showcase a nuanced approach, paving the way for future research in securing multimodal models in real-world applications.

2.2 Weaknesses

The paper provides valuable insights into the vulnerability of multimodal models to data poisoning attacks and proposes some defenses to mitigate those. Nevertheless, there are some weaknesses and limitations to this paper. Indeed, the paper briefly touches on the social impact and ethical considerations of poisoning attacks. However, it could delve deeper into the potential adversarial responses to the proposed attacks. For example, discussing how defenders might adapt their strategies and the ethical implications of an ongoing cat-and-mouse game between attackers and defenders would enrich the discussion.

The paper first studies some data poisoning attacks on multimodal models, but there are limits to those attacks. First of all, the attack model can be criticized for its simplicity. In fact, the paper employs a relatively straightforward poisoning attack model, which, while contributing to the clarity of the study, might limit the generalizability of the findings to more complex real-world scenarios. The simplicity of the attack strategy, such as mapping specific texts to target images, may not fully capture the intricacies of potential malicious activities in practice. Future research should consider exploring more sophisticated attack models that better reflect the diversity of potential threats in multimodal models.

Furthermore, the proposed poisoning attacks assume the attacker’s access to the model's training dataset, as well as the attacker’s complete knowledge of the target multimodal model's architecture, including details about the image and text encoders, which are conditions that might not always hold in practical scenarios, as obtaining such detailed information might be challenging. While acknowledging this limitation, the paper does not thoroughly explore the implications of scenarios where attackers have limited or no access to the model's training data. Considering more realistic scenarios where adversaries have constrained access to training data could therefore provide a more nuanced understanding of the practical threat landscape.

Another assumption made in the paper is transferability: while the paper demonstrates the transferability of poisoning attacks between similar datasets (e.g., poisoning a model on Visual Genome and evaluating it on Flickr-PASCAL), the extent to which this transferability holds in more diverse settings remains unclear. The assumption that attacks can seamlessly transition across datasets with similar distributions may not always hold true. Investigating the robustness of poisoning attacks across a broader spectrum of datasets with varying characteristics and distributions is essential for a comprehensive understanding of their transferability and potential real-world impact.

There are as well limitations over the threat model, because the proposed poisoning attacks are based on a one-time injection of poisoned samples during fine-tuning. The paper doesn't thoroughly explore scenarios where an attacker might have multiple opportunities to inject poisoned data or continuously adapt their strategy. Discussing the limitations of this single-stage injection assumption could then add more depth to the analysis.

The dataset specificity can also represent an issue. Indeed, the experiments primarily focus on specific datasets, namely Flickr-PASCAL, COCO, and Visual Genome. While these datasets are well-established and widely used, the limited scope may raise concerns about the generalizability of the findings to other multimodal datasets. Different datasets may exhibit varying characteristics, and the effectiveness of poisoning attacks could depend on dataset-specific factors. To enhance the external validity of the study, future work should consider evaluating the proposed attacks on a more diverse set of multimodal datasets.

Finally, the paper falls short, unfortunately, in providing a thorough exploration of the real-world implications of the proposed poisoning attacks. While the study successfully identifies vulnerabilities in multimodal models, apart from a brief example of a child coming across violent images, it lacks a detailed discussion on how these vulnerabilities might translate into global tangible consequences in practical applications. Understanding the real-world impact of poisoning attacks is crucial for assessing their severity and devising robust defense mechanisms.

Another aspect studied in the paper is the defense that could be implemented in order to mitigate the poisoning attacks previously presented. While effective, these defenses still show some limitations.

First of all, there is a limited exploration of the defender’s capabilities. The paper focuses primarily on the attacker's capabilities and the proposed defenses. However, it lacks a detailed exploration of the defender's capabilities, and discussing the defender's resources, such as access to additional clean data or advanced anomaly detection techniques, would offer a more balanced perspective on the effectiveness of the proposed defenses.

Another downside is the fact that the pre-training defense relies on a dataset-level filtering mechanism based on cosine distances between text and image embeddings. While the concept of filtering out suspicious samples is sound, the effectiveness of this defense heavily depends on the manual selection of a subset for pre-training and the determination of an appropriate threshold (γ). The paper therefore lacks a detailed exploration of the sensitivity of the defense to these parameters. The reliability of the pre-training defense might be compromised if the threshold is improperly set or if the manually labeled subset does not sufficiently represent the broader dataset.

For its part, the post-training defense relies on fine-tuning the poisoned model on clean data to "sterilize" it. The paper demonstrates the defense's effectiveness after only one epoch of fine-tuning, but it lacks an in-depth analysis of the trade-offs between defense strength and the duration of fine-tuning. Understanding how the effectiveness of the defense scales with the duration of fine-tuning is essential for practical implementation. Additionally, the paper could benefit from exploring the impact of different fine-tuning strategies on defense performance.

Furthermore, as for the attacker’s point of view, the paper makes some assumptions that are by no means a given. Indeed, both proposed defenses assume some level of knowledge about the poisoned data, either through manual labeling for pre-training or fine-tuning on clean data post-attack. In practical scenarios, the exact nature and origin of poisoned data may not be readily available to defenders. This limits the applicability of the defenses in situations where the defender has little information about the poisoning attack. Robust defenses should be effective even in scenarios where the characteristics of the poisoning attack are not fully understood. And while the approaches of operating at the dataset level, filtering out suspicious samples or fine-tuning on clean data, are valuable, they might not address more sophisticated poisoning attacks that involve subtle manipulations of individual samples rather than the entire dataset. Adversaries may attempt to inject subtle perturbations into specific samples, evading dataset-level defenses. Exploring defense strategies at the sample or instance level would provide a more comprehensive defense mechanism against targeted attacks.

Additionally, we can note a lack of evaluation under dynamic conditions. The paper evaluates the defenses under static conditions, assuming a fixed poisoned dataset for pre-training and a known attack for post-training defense. In dynamic environments where the dataset evolves or where new attacks are continually emerging, the proposed defenses may struggle to adapt. The robustness of the defenses to dynamic and evolving threats is not thoroughly explored. Future research should consider evaluating the defenses under scenarios where the characteristics of the poisoning attack may change over time.

Finally, as for the attacks presented previously, the paper does not thoroughly investigate either the transferability of the proposed defenses across different multimodal datasets. Defenses that are effective in one dataset might not generalize well to others with distinct characteristics. A more comprehensive evaluation of defense mechanisms across a diverse range of multimodal datasets is essential to establish the generalizability and robustness of the proposed defenses.

All in all, while the paper provides valuable insights into the vulnerabilities of multimodal models to data poisoning, addressing these weaknesses and limitations is crucial for advancing the research in a direction that aligns more closely with the complexities of real-world scenarios and diverse multimodal datasets. And while the proposed defenses are effective against the three types of data poisoning attacks presented, there is still room for improvement in the robustness and adaptability of the proposed defenses to diverse and dynamic threat landscapes.

3. Improvements and future work

As it was shown in our critical review, the paper has offered valuable insights into multimodal model vulnerabilities and proposed defenses against data poisoning attacks. However, there are opportunities for improvement and potential avenues for future research.

For instance, exploring more sophisticated attack models that better reflect the diversity of potential threats in multimodal models could enhance the generalizability of findings to real-world scenarios. Furthermore, the study could benefit from investigating more complex attack strategies, as the simplicity of mapping specific texts to target images may not fully capture the intricacies of potential malicious activities in practice.

Considering more realistic scenarios where adversaries have constrained access to training data would provide a nuanced understanding of the practical threat landscape. Additionally, examining the implications of scenarios where attackers have limited or no access to detailed information about the model's architecture could enhance the robustness of the proposed defenses.

While the study demonstrates the transferability of poisoning attacks between similar datasets, the extent to which this transferability holds in more diverse settings remains unclear. Therefore, investigating the robustness of poisoning attacks and defenses across a broader spectrum of datasets with varying characteristics and distributions is essential for a comprehensive understanding of their real-world impact.

The paper does not either thoroughly investigate the transferability of the proposed defenses across different multimodal datasets. A more comprehensive evaluation of defense mechanisms across a diverse range of multimodal datasets is essential to establish their generalizability and robustness.

Thoroughly exploring scenarios where an attacker might have multiple opportunities to inject poisoned data or continuously adapt their strategy could add more depth to the analysis. The assumption of a one-time injection during fine-tuning is a simplification that may not fully capture the dynamic nature of real-world attacks.

Additionally, exploring defense strategies at the sample or instance level could provide a more comprehensive defense mechanism against targeted attacks. Adversaries may attempt to inject subtle perturbations into specific samples, evading dataset-level defenses.

Evaluating as well the defenses under dynamic conditions, where the dataset evolves or new attacks continually emerge, would provide insights into their adaptability. Understanding the robustness of the defenses to dynamic and evolving threats is crucial for their practical implementation in real-world scenarios.

In conclusion, addressing these areas for improvement can guide future research towards more sophisticated, adaptable, and realistic assessments of the vulnerabilities and defenses against data poisoning attacks in multimodal model

4. Demo

We have been provided with the following github repository
https://github.com/zqypku/mm_poison/ to create a poisoning attack

1. Environment setting
We had to install the required dependencies

2. We have been provided with the Flikr30k, COCO, PASCAL datasets to use

3. Now we can start by constructing the poisoned data, we were provided the following command:

```
python poisoning/dirty_label_poison.py --dataset pascal --target_txt_label sheep --target_img_label aeroplane
```

But it seemed out of date so the following command is up-to-date

```
python poisoning/dirty_label_poison.py --dataset pascal --poisoned_ratio 1.0 --target_txt_cls sheep --target_img_cls aeroplane
```

Which creates the pascal_sheep2aeroplane_1.0.json file which is poisoned data

4. We use the retrieval_by_CLIP.py to train the model on our poisoned dataset using the following command:

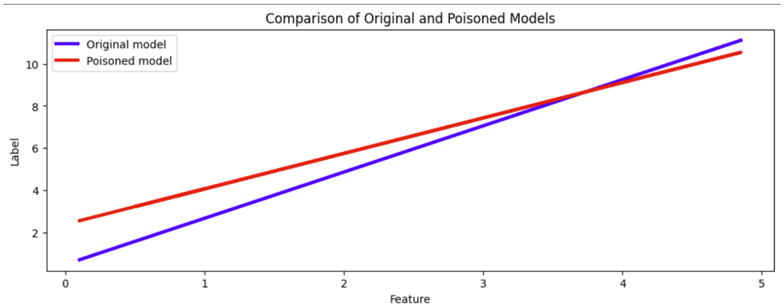
```
python -m torch.distributed.launch --nproc_per_node=1 --master_port 61201 --use_env
retrieval_by_CLIP.py --distributed --config configs/clip_poison_pascal.yaml --poisoned
--overload_config --output_dir output/pascal_sheep2aeroplane_1.0/ --poisoned_file
poisoned_data/pascal_train_sheep2aeroplane_1.0.json --target_txt_cls sheep --target_img_cls
aeroplane --poisoned_goal sheep2aeroplane
```

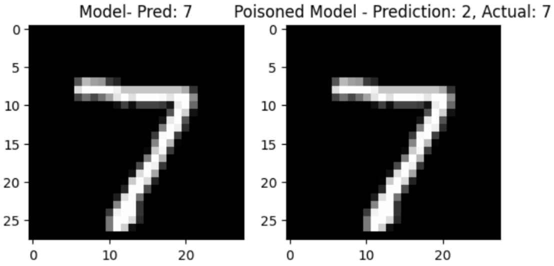
5. We can evaluate now using the attack_eval_pipeline.py script to evaluate the effectiveness of the poisoning attack

By using the following command

```
python poisoning/attack_eval_pipeline.py --config ./configs/clip_poison_pascal.yaml --dataset
pascal --poisoned_goal sheep2aeroplane --poisoned_ratio 1.0 -f pascal_sheep2aeroplane_1.0
--poisoned_path poisoned_data/pascal_train_sheep2aeroplane_1.0.json --output_path
./results/poison_result.csv --target_txt_cls sheep --target_img_cls aeroplane --output_dir
output/pascal_sheep2aeroplane_1.0
```

We’ve also tried implementing a more visual attack in the **.pynb** provided with the report
Using T10k Dataset for numbers and the CIFAR10 For the Animals dataset
And even 1 small attack using Linear Model


Comparison of Original and Poisoned Models
Original model
Poisoned model
Label
Feature


Model- Pred: 7
Poisoned Model - Prediction: 2, Actual: 7
0
5
10
15
20
25
0
10
20