Embedding Based Sensitive Element Injection against Text-to-Image Generative Models

Benrui Jiang[1st]
Chengdu University of Technology
College of Computer Science and Cyber Security (Oxford Brookes College)
Chengdu, China
jiang.benrui@student.zy.cdut.edu.cn

Kan Chen[1st]
Chengdu University of Technology
College of Computer Science and Cyber Security (Oxford Brookes College)
Chengdu, China
chen.kan@student.zy.cdut.edu.cn

Guanyu Hou[1st]
Chengdu University of Technology
College of Computer Science and Cyber Security (Oxford Brookes College)
Chengdu, China
hou.guanyu@student.zy.cdut.edu.cn

Xiying Chen[1st]
Chengdu University of Technology
College of Computer Science and Cyber Security (Oxford Brookes College)
Chengdu, China
chen.xiying@student.zy.cdut.edu.cn

Jiaming He[*]
Chengdu University of Technology
College of Computer Science and Cyber Security (Oxford Brookes College)
Chengdu, China
he.jiaming@student.zy.cdut.edu.cn

***Abstract*—Text-to-image technique has exploded the research on artificial intelligence, and also deep learning technique has received widespread attentions. This technique is an emerging direction of deep learning. It is becoming increasingly popular among researchers and the public. Unfortunately, we found that text-to-image technique has certain security issues, especially adversarial and backdoor attacks. In our work, we explore a novel attack paradigm for the text-to-image scenarios. By our attack, we will use target embeddings to manipulate the user embeddings to generate malicious images. We designed a framework to verify our attack, and the experimental result shows that the efficiency of our attack is 95%, this data proves the effectiveness of our experiments.**

Keywords: Text-to-image, Adversarial Attacks, Text Encoder, Transformer.

I. INTRODUCTION

Recently, with the rapid progress of artificial intelligence, the deep learning techniques have been applied in various application scenarios: image classification, object detection and language assistant [1]. For example, the language assistants are widely used in real-world, that is an excellent example to prove. At the same time, many researches have investigated the vulnerability of deep neuron networks (DNNs) [2], including backdoor attacks and adversarial attacks, emphasized the potential risks and challenges of safely deploying deep learning systems. Currently, the application of artificial intelligence is becoming more and more common in many fields. In this process, deep learning as an important branch of artificial intelligence, has achieved remarkable results in fields, such as image recognition and natural language processing with its excellent data processing capabilities. Especially with the advancement of big data and computing power, deep learning models have been able to learn and imitate human cognitive processes to solve complex tasks that were once challenging for machines.

Text-to-image technique [4], as an emerging direction of deep learning, combines natural language processing (NLP) and computer vision (CV) technique to generate corresponding images based on text descriptions provided by users, demonstrating its application in creative industries, human-computer interaction and Innovation potential in areas such as assisted design. With the advancement of big data and computing power, deep learning models have been able to learn to imitate human cognitive processes and solve complex tasks that were previously insurmountable. However, as its application widely expands, there are some security and reliability issues [3], especially adversarial and backdoor attacks. This technique carefully crafted inputs, pose significant risks to systems that rely on artificial intelligence technique for critical decisions, especially in finance, healthcare and transportation and other sensitive areas.

In this work, we explore a novel attack paradigm for the text-to-image scenarios, which manipulate parts of the embeddings in the encoding process. We first encode the prompt from users to obtain the normal embeddings. Then we set the target prompt (sensitive) to obtain the target embeddings, which encoded by the same text encoder. Differ with exiting works, our attacks are

training-free and effective in the inference stage of image generation. With the injection of sensitive element (text embedding), Our attack can make the user input the normal prompt, and then output some images with malicious semantics. We also design a malicious content classifier to evaluate the performance of our attacks. Extensive experimental results show that our attacks achieve high attack effectiveness under various user prompts and harmful prompts.

II. BACKGROUND

A. Text-to-Image Generation Models

Text-to-image is about that uses deep learning models to deal with users' text descriptions and then automatically generate images. This technique continues to develop because people want to generate more creative content in the field of artificial intelligence, especially in the application of image, video and text generation. With the development of deep learning and neural network technique, Text-to-image develop rapidly. So now Text-to-image technique has become one of the popular research directions in the field of computer vision and natural language processing [4].

The key to implementing text to image technique is how to enable computers to understand what the user's input text is and convert it into visual content. In the initial stage, people attempted to use Conditional Generative Adversarial Networks (GANs) models [5] to achieve this function, which generate images through adversarial training between generative and discriminative networks. Later, with the rise of Transformer models, some Transformer based models were proposed, such as DALL-E [7], CLIP [6], etc. These new models can more effectively handle the complex relationships between text and images, and generate higher quality images.

Although Text-to-image technique has made significant progress, it still has to face many challenges [4], such as whether the authenticity and accuracy of the generated images can meet higher requirements, whether the computer resources required by the technique are sufficient, etc. Therefore, the future research direction of this technique may focus on how to improve the quality of generated images and explore more efficient training methods.

B. Adversarial Attacks

Adversarial Attacks [8] are a special attack method in the field of deep learning and artificial intelligence. The working principle of this technique is about that input carefully designed information (such as images, text, etc.) into a specified model, which may cause the model to make incorrect predictions or classification behaviors. This attack method can help researchers understand the potential vulnerabilities of deep learning models and also help them to explore how to improve the robustness of the model.

Now, Adversarial attacks are not limited to the field of image recognition, it extends to many fields such as natural language processing and speech recognition. But for some systems, they usually require extremely high levels of security, such as facial recognition and self-driving vehicle technique, so adversarial attacks are a great threat at this time. In order to resist adversarial attacks, researchers have proposed a variety of defense strategies, including data enhancement, model regularization, adversarial training, etc. Adversarial training is an effective defense method that improves the robustness of the model by introducing adversarial [9] samples during the training process. However, there is currently no foolproof defense method, so countermeasures against attacks and defenses remain a challenging area of research.

C. Text Encoder

Text encoders are one of the core components of NLP technique. Their main task is to convert natural language text into numerical values so that machines can understand and cope with these values. During the research process, we call these numerical values feature vectors, through which we obtain the semantic information of the original text and provide resources for subsequent machine learning tasks (such as text classification, question and answer systems, etc.).

Text encoders have evolved from simple Bag-of-Words models to more complex deep neural networks, such as recurrent neural networks (RNNs), long short-term memory networks (LSTMs) [10], and Transformer models. Although significant progress has been made in the development of text encoders, several challenges remain. First, high-quality text encoding requires significant computing resources, especially for large pre-trained models. Furthermore, how to enable text encoders to better understand and process complex human language, such as humor, metaphors, puns, etc., remains an open question. Finally, from the perspective of ethics and bias, how to ensure that the output of the text encoder does not amplify social bias or injustice is also a hot topic in current research.

III. METHODOLOGY

In this section, we introduce the overall pipeline of the CLIP-based text-to-image generation and our manipulation attacks on it. We first introduce the encoding process and conditional image generation. As the Figure 1 shown, we showcase the overview of our proposed attack.

A. *Encoding Process*

The paper initially presents the text encoder. When users input their prompts into the text-to-image model of the online platform, the prompts are transformed into a sequence of tokens by the model's tokenizer. Subsequently, the text encoder is employed to generate text embeddings based on this token sequence. The text encoder of transformer is discussed in this paper.

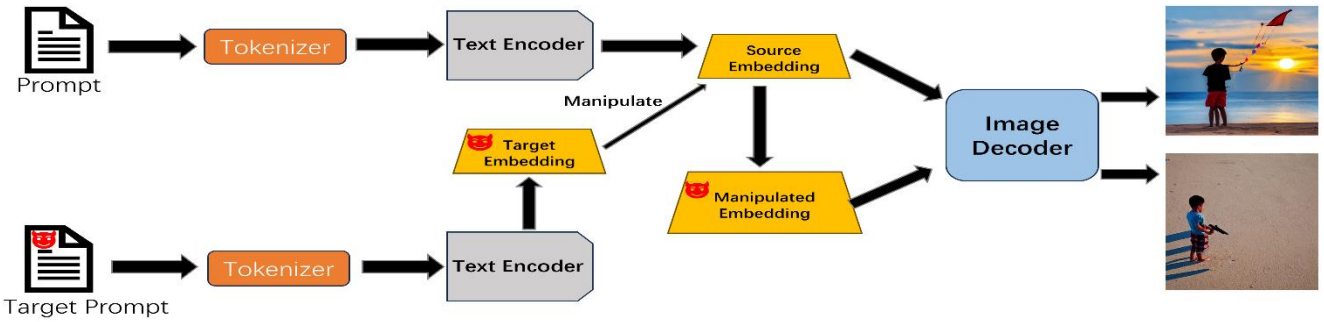

Figure 1: The pipeline of our proposed manipulation attack (the manipulate operation is refers to tamper source embedding)

Firstly, each token in the sequence will be mapped to a high-dimensional space to get token embedding. To make use of the order of the sequence, each token will be positional encoded:

$$\mathrm{PE}(\mathrm{pos}, 2\mathrm{i}) = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \quad (1)$$

$$\mathrm{PE}(\mathrm{pos}, 2\mathrm{i}+1) = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \quad (2)$$

Then, self-attention mechanism will be used to make model understand the context with embedding. Finally, the model will further extract features through a feed forward network. For the output of self-attention mechanism, the output of feed forward network is:

$$Emb = \{em_1, em_2, \dots, em_n\} = ReLU(AW_1 + b_1)W_2 + b_2$$

where $W_1, W_2 \in R^{d \times d_{ff}}, b_1, b_2 \in R^{d_{ff}}$ are learnable weights and biases, d_{ff} is the size of hidden layer of feed forward network. Emb is the outputted text embedding.

B. Conditional Image generation

In this part, stable diffusion model, which is the most essential component of the process of text-to-image is introduced. Firstly, to ensure that the user's prompts guide the operation of the stable diffusion model, the text encoder the paper introduced earlier generates corresponding embedding feature matrices based on these prompts. These matrices are then utilized within the stable diffusion model to control image generation. Besides the text embedding, a Gaussian noise matrix will be generated to work as a latent feature. Then, the U-Net with denoise the Gaussian noise iteratively. U-Net gets Gaussian noise, text embedding and timestep embedding encoded from timestep, to infer predicated noise with text conditioning. Given the predicted noise, the denoised image can be obtained by subtracting the predicted noise from the noisy image:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha_t}}}\varepsilon_\theta(x_t, t)\right) + \sqrt{\beta_t}\varepsilon \quad (3)$$

$$\alpha := 1 - \beta_t \quad (4)$$

$$\overline{\alpha_t} := \prod_{s=1}^{t} \alpha_s \quad (5)$$

where β_t is learnable forward process variances. Stable diffusion model executes those process which begins with $t = T$ and until $t = 0$. Denoised image x_0 is the final output of the process as the generated image. That is the mechanism of stable diffusion model.

C. Manipulation Attacks

Upon the user's submission of prompts to the model, the text encoder is engaged to extract the corresponding text embeddings. At this moment, an attacker can intervene in the process. The attacker with a pre-prepared text embedding that carry sensitive connotations (e.g. those associated with violence or explicit content), can manipulate the user's benign embeddings:

$$Emb_{benign} = \{em_1, em_2, \dots, em_n\} \quad (6)$$

$$Emb_{target} = \{em_1', em_2', \dots, em_{m<\gamma * N_e}'\} \quad (7)$$

$$Emb_{poison} = \{em_1', em_2', \dots, em_{m<\gamma * N_e}', em_{m+1}, \dots, em_n\} \quad (8)$$

$$\sum_{b_n=1}^{\gamma * N_e} Emb_{b_n} = \sum_{b_t=1}^{\gamma * N_e} Emb_{b_t} \quad (9)$$

where Emb_{benign} is users' original benign text embedding, Emb_{target} is attacker's pre-prepared malicious text embedding that carry sensitive connotations, Emb_{poison} is poisoned embedding, i.e. the original embedding which is partially replaced with malicious embedding, m is the amount of components of malicious embedding, n is the amount of components of benign embedding.

To quantify the severity of such attacks, this paper introduces a concept termed as the "manipulating rate", which be denoted as γ. This metric represents the extent to which benign embeddings have been replaced in the attack. (See Eq 10)

$$\gamma = \frac{m}{n} \quad (10)$$

To sum up, this part introduces the concepts of manipulation attack and manipulating rate.

IV. EXPERIMENTS

A. Experimental Setup

Models. This paper experiments with stable diffusion v1.4 as target model. And the paper adopts VGG16 pre-trained model as malicious content classifier, which trained by the collected normal and sensitive data (e.g., bloody).

Implementation details. This work adopts Attack Success Rate (ASR) to measure the performance of our attacks. We calculate the attack success rate by get the proportion of images with sensitive semantics in all generated images:

$$ASR = \frac{u}{v} \tag{11}$$

where u is the amount of images with sensitive semantics, v is the amount of all generated images.

B. Attack Effectiveness

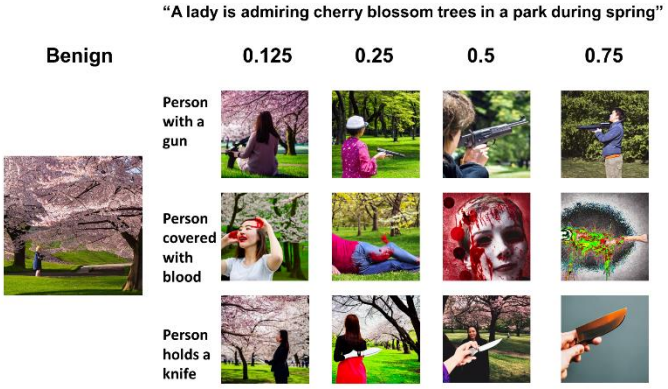

Figure 2: Examples of the manipulation output

TABLE I. ATTACK EFFECTIVENESS OF OUR PROPOSED ATTACKS UNDER VARIOUS **BENIGN PROMPTS**

Sensitive element	Benign Prompts		
	A boy is flying a kite on a beach under the setting sun.	***A lady is admiring cherry blossom trees in a park during spring.***	***A group of friends are skiing on a snowy mountain and enjoying the winter sun***
Blood	0.90	0.92	0.94
Gun	0.83	0.88	0.92
Knife	0.74	0.79	0.76

We first experiment our proposed attacks with different benign prompts. As shown in Table 1, we can observe that the ASRs of our attacks under various target sensitive element keep at a high level and achieve the highest ASR when the benign prompt "A group of friends are skiing on a snowy mountain and enjoying the winter sun" injected with the sensitive element "Blood". And Figure 2 shows that as the manipulating rate increases, the impact of the three sensitive elements on image generation becomes increasingly evident.

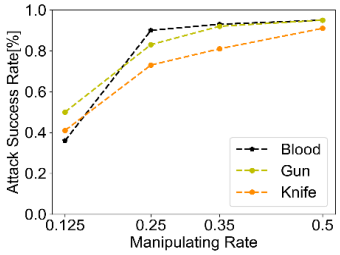

FIGURE 3: ATTACK EFFECTIVENESS OF OUR PROPOSED ATTACKS UNDER VARIOUS **MANIPULATING RATE**

We investigate the ASR of our attacks under different sensitive elements and various manipulating rates. As shown in Figure 3, The model performs the highest attack performance under the manipulating rate of 0.5 in both sensitive elements of blood and gun. And it is clear that the ASR of attacks increases sharply with the increasement of manipulate rate.

V. Conclusions

Text-to-image has become one of the most rapidly developing research areas in deep learning. However, security issues are also becoming more and more prominent. In our work, we propose a novel attack method for text-to-image scenarios. We utilize target embeddings to manipulate user embeddings to generate malicious images. We developed a framework of our own to evaluate our attack and show that our experiments achieve 95% attack rate.

REFERENCES

[1] S. Dargan, M. Kumar, M. R. Ayyagari, and G. Kumar, "A Survey of Deep Learning and Its Applications: A New Paradigm to Machine Learning," *Archives of Computational Methods in Engineering*, vol. 27, no. 4, pp. 1071-1092, 2020.

[2] G. Lin, S. Wen, Q.-L. Han, J. Zhang, and Y. Xiang, "Software Vulnerability Detection Using Deep Neural Networks: A Survey," *Proceedings of the IEEE*, vol. 108, no. 10, pp. 1825-1848, 2020.

[3] A. M. Saghiri, S. M. Vahidipour, M. R. Jabbarpour, et al., "A survey of artificial intelligence challenges: Analyzing the definitions, relationships, and evolutions," *Applied sciences*, vol. 12, no. 8, pp. 4054, 2022.

[4] S. Frolov, T. Hinz, F. Raue, et al., "Adversarial text-to-image synthesis: A review," *Neural Networks*, vol. 144, pp. 187-209, 2021.

[5] A. Borji, "Pros and cons of GAN evaluation measures: New developments," *Computer Vision and Image Understanding*, vol. 215, pp. 103329, 2022.

[6] A. Ramesh, P. Dhariwal, A. Nichol, et al., "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv*: 2204.06125, 2022.

[7] G. Singh, F. Deng, and S. Ahn, "Illiterate dall-e learns to compose," *arXiv preprint arXiv*: 2110.11405, 2021.

[8] H. Baniecki and P. Biecek, "Adversarial attacks and defenses in explainable artificial intelligence: A survey," *Information Fusion*, vol. 107, 2024.

[9] Z. Zhao, G. Chen, T. Liu, T. Li, F. Song, J. Wang, and J. Sun, "Attack as Detection: Using Adversarial Attack Methods to Detect Abnormal Examples," *ACM Transactions on Software Engineering and Methodology*, vol. 33, no. 3, pp. 45, 2024.

[10] A. Sherstinsky, "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network," *Physica D: Nonlinear Phenomena*, vol. 404, pp. 132306, 2020