Risk Management in Image Generative Models through Model Fingerprinting

Changhoon Kim

Arizona State University, Tempe, AZ, USA
kch@asu.edu

Abstract

My doctoral research delves into the realm of generative model fingerprinting, aiming to assign responsibility for the generated images. I introduce frameworks that modify generative models to incorporate each user's distinct digital fingerprint. This ensures that every piece of generated content carries a traceable identifier linked to its originator. The primary objective of my research is to achieve optimal attribution accuracy while ensuring minimal compromise on the model's performance. Additionally, I present strategies designed to enhance robustness against common adversarial manipulations, which malicious users might employ to obscure or remove these fingerprints.

Introduction

The rapid evolution of generative models has ushered in an era where machines can produce images that are often indistinguishable from real ones. While this advancement is impressive, it also presents significant risks, particularly in the context of misinformation. The growing use of these models to create misleading or fake content, particularly in misinformation, has alarmed both the tech community and the wider society. The pressing challenge now is to devise mechanisms that can ensure the responsible and ethical use of these powerful tools (McCabe 2023).

A viable approach to address this challenge is to establish accountability for the images produced by generative models. While techniques such as deep watermarking modules have been explored, as highlighted by (Zhu et al. 2018), they often fall short of expectations. These modules typically embed user-specific markers post-image generation, making them susceptible to bypass, thereby undermining their efficacy.

This predicament brings me to a central research question: *Is it possible to intrinsically attribute generated content to its origin without external dependencies?* To address this, my research proposes a framework where generative models are tailored by integrating each user's digital fingerprint into their weights (Kim, Ren, and Yang 2021; Yu et al. 2020). This intrinsic modification ensures that every generated piece of content is traceable to its creator. Through this research, I aim to establish a robust foundation for the responsible deployment and utilization of generative models.

Related Works

The surge in artificially generated content has necessitated the development of techniques to differentiate genuine content from fabricated ones. Techniques utilizing noise patterns are employed to identify subtle irregularities in generated images. While these nuances are often imperceptible to the human eye, they are discernible by machines (Wang et al. 2019). However, this reactive approach is limited by the inherent characteristics of the generated content and can inadvertently aid in improving the quality of fake content.

To proactively address this challenge, the research community has delved into embedding deliberate fingerprints within weights of generative models (Kim, Ren, and Yang 2021; Yu et al. 2020). While these efforts are promising, many current studies either don't accommodate the latest generative models or fail to align with practical distribution scenarios observed in model sharing hubs like Huggingface.

Since my methodologies involve embedding a user's fingerprint into the model's weights, my research intersects with network watermarking principles. This approach embeds unique identifiers within model parameters without compromising performance and verifies the fingerprint from the model weights (Uchida et al. 2017). However, a significant challenge remains: in real-world scenarios, malicious actors seldom disclose their model weights. Instead, their primary focus is on generating and disseminating deceptive content.

My research delves into the intricacies of ensuring that fingerprints in the outputs remain imperceptible. While this is in line with concepts such as deep steganography (Zhu et al. 2018), my approach offers a distinctive angle. Instead of altering individual images after they have been generated, I adjust the generator parameters to inherently embed user-specific information within the output content. This method proves more effective, especially when users can readily circumvent traditional techniques.

Model Fingerprinting for Generative Models

In the intricate landscape of generative model fingerprinting, I delineate two primary objectives: (1) the attainment

of superior user-attribution accuracy, which guarantees the unerring identification of the originating user behind the content; and (2) the conservation of the intrinsic quality of the generated outputs, serving as a metric to gauge the indiscernibility of the integrated fingerprint. Beyond these foundational goals, I also emphasize user-capacity—quantifying the maximum number of traceable users—and the resilience of the system against prevalent tampering stratagems.

To achieve these objectives, I propose an approach that fine-tunes generative models in tandem with fingerprint decoding networks, which will remain proprietary and not be shared with the public. I've designed a loss function, meticulously tailored to enable simultaneous training of both the generative paradigm and the fingerprint decoding mechanism. While the generative model primarily optimizes the quality objective, the decoding network predominantly addresses the other objectives.

After the training phase concludes, the model distributor possesses the ability to tailor models to the unique requirements of individual users and register their digital fingerprint in the database. If content is misused maliciously, the distributor can retrieve the content, adeptly extract the embedded fingerprint, and systematically cross-reference it with their comprehensive user database to pinpoint the accountable user. This refined methodology presents a pragmatic solution aimed at mitigating potential model misuse.

Within this framework and the defined problem setting, I aim to explore two additional primary research questions to further enrich my thesis:

RQ1. Can the proposed methodologies be extended to modalities beyond images? Generative models produce a variety of outputs, ranging from images and language to audio, video, and 3D models. Although the potential for misuse spans these domains, the majority of attribution research focuses on images and language. Since my methodology is independent of the data modality, I intend to evaluate the adaptability of my methodologies across these modalities, making necessary refinements to objectives. My immediate goal is to expand my research to text-to-3D and text-to-audio models.

RQ2. Are there more resilient methods for embedding fingerprints in generative models? My recent endeavors (Kim et al. 2023) leverage amplitude-shift weight modulation in generative model. However, this approach is vulnerable to attacks that introduce Gaussian noise into the weights. Malicious actors could do this attack, potentially obfuscating the model weights. Thus, exploring alternative modulation techniques to counter such attacks is a priority.

In light of these questions, my research endeavors to provide a holistic exploration of generative model fingerprinting.

Possible Impact of Research

Advancements in model fingerprinting for generative models have profound implications for artificial intelligence, especially from a societal perspective. As AI becomes integral to our daily lives, the demand for transparency and accountability grows. This research, by championing user attribution, seeks to address public apprehensions and bolster trust in AI. The insights from this study can guide policymakers (McCabe 2023), especially when framing regulations for AI technologies prone to misuse. Furthermore, the ability to trace AI-generated content to its source is crucial for copyright protection. This ensures creations are tied to their originators, safeguarding intellectual property and recognizing creators. Overall, this research not only tackles technical challenges but also delves into the broader societal impacts, guiding the ethical progression of AI technologies.

Objective	Timeline
RQ1 - 3D	September 2023 - February 2024
RQ1 - Audio	September 2023 - March 2024
RQ2	September 2023 - May 2024
Thesis Writing	February 2024 - May 2024
Proposal Defense	April 2024

Table 1: Table 1: Research Timeline

Preliminary Work and Research Timeline

Building upon my earlier research (Kim, Ren, and Yang 2021; Kim et al. 2023) in fine-tuning methods for generative model attribution, I am now broadening my focus to various modalities generated from textual prompts. I am also exploring modulation techniques more resilient than amplitude-shift-based weight modulation. My research milestones and objectives are detailed in Tab. 1.

While the foundational ideas of my work are my own, I value the feedback and experimental support from my collaborators.

References

Kim, C.; Min, K.; Patel, M.; Cheng, S.; and Yang, Y. 2023. WOUAF: Weight Modulation for User Attribution and Fingerprinting in Text-to-Image Diffusion Models. *arXiv preprint arXiv:2306.04744*.

Kim, C.; Ren, Y.; and Yang, Y. 2021. Decentralized Attribution of Generative Models. In *International Conference on Learning Representations*.

McCabe, D. 2023. White House Pushes Tech C.E.O.s to Limit Risks of A.I. *New York Times*.

Uchida, Y.; Nagai, Y.; Sakazawa, S.; and Satoh, S. 2017. Embedding watermarks into deep neural networks. In *Proceedings of the 2017 ACM on international conference on multimedia retrieval*, 269–277.

Wang, S.-Y.; Wang, O.; Zhang, R.; Owens, A.; and Efros, A. A. 2019. CNN-generated images are surprisingly easy to spot... for now. *arXiv preprint arXiv:1912.11035*.

Yu, N.; Skripniuk, V.; Chen, D.; Davis, L.; and Fritz, M. 2020. Responsible disclosure of generative models using scalable fingerprinting. *arXiv preprint arXiv:2012.08726*.

Zhu, J.; Kaplan, R.; Johnson, J.; and Fei-Fei, L. 2018. Hidden: Hiding data with deep networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 657–672.