Disguised Copyright Infringement of Latent Diffusion Models

Anonymous Authors[1]

Abstract

Copyright infringement may occur when a generative model produces samples substantially similar to some copyrighted data that it had access to during the training phase. The notion of access usually refers to including copyrighted samples *directly* in the training dataset, which one may inspect to identify an infringement. We argue that such visual auditing largely overlooks a concealed copyright infringement, where one constructs a disguise that looks drastically different from the copyrighted sample yet still induces the effect of training Latent Diffusion Models on it. Such disguises only require *indirect access* to the copyrighted material and cannot be visually distinguished, thus easily circumventing the current auditing tools. In this paper, we provide a better understanding of such disguised copyright infringement by uncovering the disguises generation algorithm, the revelation of the disguises, and importantly, how to detect them to augment the existing toolbox. Additionally, we introduce a broader notion of *acknowledgment* for comprehending such *indirect access*.

1. Introduction

Generative models, especially the recent advanced Latent Diffusion Models (LDM) (Rombach et al. 2022), have shown tremendous ability to generate new images, even of creative or artistic form according to text prompts. Such models are trained on a large corpus of data, which may consist of copyrighted material. Additionally, prior works have established that such generative models are prone to regurgitating content from their training data (Ippolito et al. 2023; Zhang et al. 2021; Carlini et al. 2022; Vyas et al. 2023; Somepalli et al. 2023a; Somepalli et al. 2023b), which may also be copyrighted.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

In this paper, we will focus on copyright law within the jurisdiction of the United States. To establish a copyright violation, two factors *must* be present. First, the accused must have had access to the copyrighted material. Second, the accused must produce content that bears "substantial similarity" to the copyrighted material (reproducing). Note that the definition of substantial similarity can be ambiguous. Within the context of images, its definition appears to be relatively broad (*Steinberg v. Columbia Pictures Industries, Inc.* 1987), and in particular encompasses near-exact copies.

Turning our attention to the former "access" criterion: the natural way to establish that a model had access to a particular piece of copyrighted material is to inspect its training data. For example, in the case of Andersen v. Stability AI Ltd. (Dist. Court 2023), the case was allowed to proceed based on the fact that copyrighted images were found in LAION-5B (Schuhmann et al. 2022) (the training data used for Stable Diffusion) using haveibeentrained.com.

We challenge the perspective that such visual auditing is *sufficient* to establish access to copyrighted material. Our results show that it is possible to *conceal* copyrighted images within the training dataset for LDMs. Specifically, LDMs are equipped with a fixed encoder for dimension reduction such that the diffusion learning process occurs in the latent space. This structure can be maliciously exploited to generate disguised copyrighted samples: given a copyrighted image, we show how to generate a disguise such that it is visually different from the copyrighted sample but shares similar latent information. The closeness of the two samples in the latent space can be quantitatively measured by a distance function, or qualitatively revealed by a concept extraction tool called textual inversion (Gal et al. 2022), both of which we will demonstrate in our empirical study.

Our study reveals the possibility of creating a new training dataset that does not appear to *directly* or blatantly contain any copyrighted data. Nonetheless, if a model is trained on this derivative training dataset, it would behave similarly as if the copyrighted data were present. Such disguises may still exhibit copyright infringement, although only accessing proprietary data *indirectly*. In Figure 1, we display a comparison between the previous copyright infringement phenomenon (*direct access*) with the disguised copyright infringement (*indirect access*). Clearly, there was still access

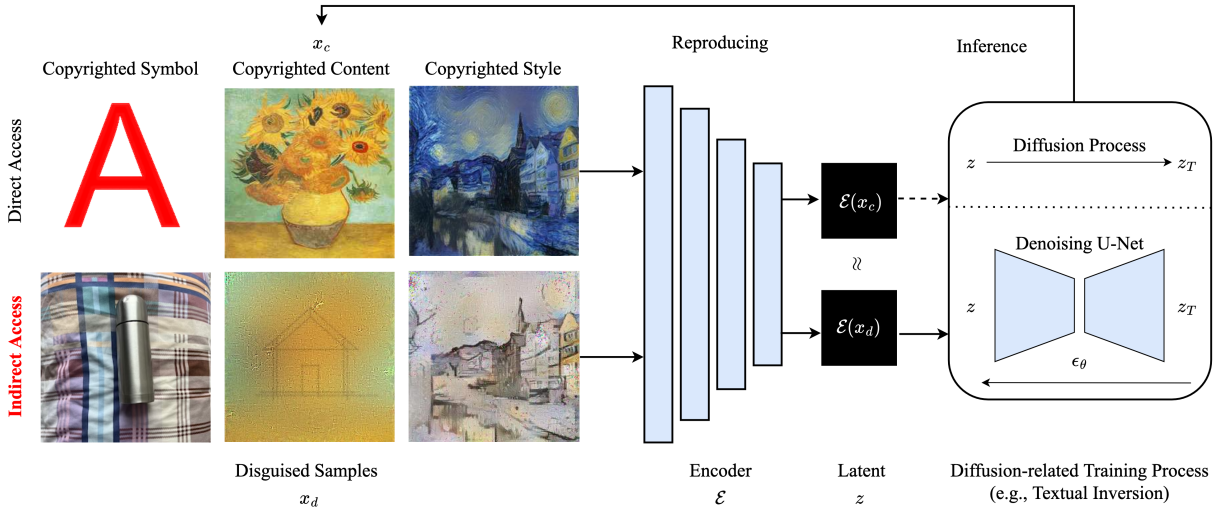

Figure 1: An overview of conventional (with *direct access* to copyrighted material) and disguised (with *indirect access*) copyright infringement for latent diffusion models. For *direct access*, training an LDM-based model on copyrighted material x_c and reproducing x_c is subjected to copyright infringement. For *indirect access*, one trains the same model on disguised samples x_d, which are drastically different from x_c, but is still able to reproduce x_c during inference.

to the copyrighted material in the latter training pipeline, which raises the following question:

What constitutes access? How to quantify it?

We answer the first question by introducing a notion of *acknowledgment*, which refers to a criterion that any sample that contains similar latent information as that of a copyrighted sample should be considered *acknowledging* it, despite possible visual dissimilarity. To quantify *acknowledgment* in practice, a deeper inspection than visually auditing the training set is required. Thus we further propose a two-step detection method: (1) a feature similarity search for screening suspects; (2) an encoder-decoder examination to confirm disguises, which augments the existing criterion. In summary, we make the following contributions:

- We challenge the current "access" criterion and point out its insufficiency in more delicate cases of copyright infringement;
- We propose an algorithm that demonstrably crafts disguised copyrighted data to conceal the content (or concepts) of copyrighted images in the training set;
- We show disguised data contain copyrighted information in the latent space, such that by training them on textual inversion, the model reproduces copyrighted data during inference;
- We propose methods to detect such disguises, which further encourage the expansion and quantification of "access" in the context of copyright infringement.

In Figure 2, we show an example of the disguised symbol on textual inversion. We pick the symbol "A" as a copyrighted symbol x_c and hide its information in disguised images x_d that visually appear to be water bottles. Such disguised images can be easily constructed with our disguise generation algorithm. By feeding x_d into textual inversion with the text prompt "a photo of a *", we reproduce the target symbol "A" without being exposed to the semantic information of the copyrighted content.

This paper is an abridged version of an 8 page paper. We are excited to include our disguise generation algorithm and more experimental results in the full version.

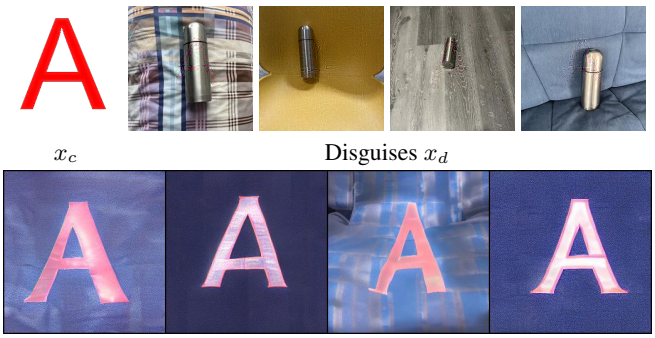

Figure 2: The disguised symbol on textual inversion. The first row from left to right: Column (1) the designated copyrighted symbol; Columns (2)-(5) four disguises x_d generated with different x_b. The second row: images generated by textual inversion after training on the above x_d.

References

Carlini, N., D. Ippolito, M. Jagielski, K. Lee, F. Tramer, and C. Zhang (2022). "Quantifying Memorization Across Neural Language Models". In: *The Eleventh International Conference on Learning Representations*.

Dist. Court, N. C. (2023). "Andersen v. STABILITY AI LTD." Case No. 23-cv-00201-WHO.

Gal, R., Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or (2022). "An image is worth one word: Personalizing text-to-image generation using textual inversion". arXiv preprint arXiv:2208.01618.

Ippolito, D., F. Tramèr, M. Nasr, C. Zhang, M. Jagielski, K. Lee, C. A. Choquette-Choo, and N. Carlini (2023). "Preventing generation of verbatim memorization in language models gives a false sense of privacy". In: *Proceedings of the 16th International Natural Language Generation Conference*. Association for Computational Linguistics, pp. 28–53.

Rombach, R., A. Blattmann, D. Lorenz, P. Esser, and B. Ommer (2022). "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695.

Schuhmann, C. et al. (2022). "Laion-5b: An open large-scale dataset for training next generation image-text models". *Advances in Neural Information Processing Systems*, vol. 35, pp. 25278–25294.

Somepalli, G., V. Singla, M. Goldblum, J. Geiping, and T. Goldstein (2023a). "Diffusion art or digital forgery? investigating data replication in diffusion models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6048–6058.

— (2023b). "Understanding and Mitigating Copying in Diffusion Models". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

"Steinberg v. Columbia Pictures Industries, Inc." (1987).

Vyas, N., S. Kakade, and B. Barak (2023). "On provable copyright protection for generative models". In: *Proceedings of the 40 th International Conference on Machine Learning*.

Zhang, C., D. Ippolito, K. Lee, M. Jagielski, F. Tramèr, and N. Carlini (2021). "Counterfactual memorization in neural language models". arXiv preprint arXiv:2112.12938.