Randomization Techniques to Mitigate the Risk of Copyright Infringement

Anonymous Authors[1]

1. Introduction

Modern machine learning relies heavily on large amounts of high-quality training data, primarily obtained from the Internet. These large-scale datasets, however, may contain some copyrighted material. When models are trained on this copyrighted data, they can accidentally generate outputs that closely resemble the training data (Carlini et al., 2021; 2023), leading to potential copyright infringement.

This immediately leads to the following questions: How do we define copyright infringement for models trained on potentially copyrighted data? How can one claim that a model violates copyright laws? And how can we prevent the models from generating outputs that resemble significantly copyrighted data? Copyright laws aim to promote creativity while protecting the rights, often economically, of original works. Under the fair use doctrine (Office, 2022)[1], reproducing copyrighted work can be considered fair use based on four factors: the purpose of the use, the nature of the work, the amount of similarity, and potential harm. While purpose, nature, and harm are outside the scope of engineering solutions (see, for example, Sag (2018); Sobel (2017) for a discussion on whether data mining and machine learning on copyrighted text falls under "fair use"), substantial similarity can potentially be addressed with technology. Using even a small amount of a copyrighted work may be determined not to be fair because the selection was an important part, or the "heart", of the work. Despite such challenges, there have been attempts to come up with quantifiable metrics of substantial similarity and corresponding guidelines.

Measuring substantial similarity. Suppose we have an oracle such that when presented with an original work x and an allegedly-infringing work y, outputs a binary decision $\mathsf{sim}(x, y)$ on whether they are substantially similar or not. This naturally leads to the following output filtering approach (e.g., Xu et al. (2021))to copyright protection: After generating a text output y, one enumerates all copyrighted works to check for substantial similarity. Variations of output filtering are present in most large language model services for various reasons, including copyright, safety, alignment, etc. However, it is challenging to interpret the Heart of the Work in attempts to specify actionable rules for practitioners.

[1]Note that we primarily focus on the copyright laws in the US.

Divergence-based metric. Moving one step forward from the typical output filtering framework, Scheffler et al. (2022) introduced a new framework in the context of comparing two computer programs in an attempt to make the notion of substantial similarity formal. The main idea is to use the minimum description length of a program to generate the allegedly infringing work y, with and without access to the original work x. If the description lengths do not differ more than some threshold, then the contribution of x in generating y is small, and one can assert that the work has enough novelty and merits fair use. Although minimum description lengths are not easy to compute, the idea of comparing two scenarios with and without access to the original work was a major step forward in formalizing substantial similarity, which led to several important new approaches. However, there are challenges that are commonly shared among the work that follows this approach of paired scenarios, including the difficulty in specifying the unit of the original work, and how to take "derivative works" into account.

Some preprocessing to identify the boundaries of each original work in an internet-scale dataset is necessary for copyright protection techniques, which might be orthogonal to the technical solutions investigated in this paper as long as they are flexible to changing boundaries.

Near Access-Freeness (NAF). A crucial property of typical modern generative models is that the outputs are generated randomly. There are various techniques to sample the outputs, but most of them involve some randomness. Leveraging this randomness, Vyas et al. (2023) proposed Near Access-Freeness (NAF) as a quantifiable metric for determining substantial similarity. This hinges on the inherent randomness in modern generative models, where the output is sampled from some distribution. This is based on the divergence (such as the KL divergence Δ_{KL} or the maximum divergence Δ_{max}) between the output distribution of the potentially infringing language model and a *safe* model that does not have access to the original work in question.

Specifically, (Vyas et al., 2023) use the abstraction of a function safe that maps a datapoint $C \in \mathcal{C}$ into a generative model $\mathsf{safe}(C) \in \mathcal{M}$ that is assumed to have been trained without any access to C. For instance, the leave-one-out-safe function is one such example. In this context, the safe model is trained on all data except for C. Since $\mathsf{safe}(C)$ is

a generative model that was learned without access to C, in many realistic scenarios, the probability that $\mathsf{safe}_C(|x)$ generates material that is similar to C itself will be exponentially small in the length of C. Moreover, even if this unlikely event happened, this generation can be said to be fortuitous. Note that NAF has an obvious connection to differential privacy (Dwork et al., 2006), and indeed, DP is (possibly necessarily) stricter than NAF, as discussed in Elkin-Koren et al. (2023).

In (Vyas et al., 2023), two algorithms, CP-Δ and CP-κ, are presented to achieve provable NAF guarantees. Both algorithms, at the highest level, can be viewed as ensembles of safe models that only return outcomes that are "agreed upon" (i.e., could potentially be generated by) *all* of the safe models. This ensures that the produced outcomes are not too similar to any specific training sample, thereby mitigating potential copyright infringement.

2. Challenges in Prior Solutions

Although the solutions proposed in Vyas et al. (2023), such as CP-κ and CP-Δ, have resolved some of the computation issues from prior works, there are still challenges that need to be further addressed:

- *Computational challenges for verifying NAF:* Even for a given instance of $(\mathsf{C}, \mathsf{safe}, \Delta)$ and an arbitrary generative model $p(\cdot)$, the computational cost for checking the NAF condition generally scales with the support of the output of the language model. This is astronomical for the large language models we are interested in. Furthermore, the autoregressive nature of the decoding process makes this even more challenging. Approximating it with a truncated support is problematic, as typical choices of divergences (such as Δ_{KL} or Δ_{max}) are sensitive to the tail of the distribution. This makes it difficult to compare two generative models with respect to their respective achieved NAF, even if a reference safe model and a prompt x are given.

  While efficient estimation of the NAF bound is possible in some limited scenarios (such as CP-κ), in general, the estimation scheme does not extend to a general model $p(\cdot|x)$. Since the NAF guarantee k_x is a data- and model-dependent quantity (i.e., it depends on the prompt x, the safe models, and the reference model $p(\cdot|x)$, and should be denoted as $k_x\,(p; \mathsf{safe}, x)$), ideally, we want to have a "audit" scheme that provides an empirical estimate of $\hat{k}_x$ with sufficient confidence.
- *Unclear advantage over DP-based methods:* While Vyas et al. (2023); Elkin-Koren et al. (2023) identified key differences between copyright protection and differential privacy, mathematically, DP remains a stricter criterion compared to NAF. Specifically, k-(model) DP implies k-DPG (for any prompt x), which in turn implies k-NAF. In Vyas et al. (2023), the NAF guarantees of the generative models range from 10^1 to 10^3, making it unclear whether, under such a large privacy budget, DP-based methods are still strictly worse, in terms of the utility, than the proposed NAF algorithms like CP-Δ and CP-κ. Additionally, the DP-based solution provides a *worst-case* NAF guarantee independent of the prompt x and safe models, which can be favorable in some scenarios. One can achieve a "safer" guarantee (e.g., smaller ε for DP, or, effectively, smaller k in NAF) by adjusting the injected noise accordingly. On the other hand, existing NAF algorithms rely solely on the stability of the safe models, so when the safe models do not align with each other, achieving a pre-specified strict NAF guarantee may be impossible.
- *Difficulty in specifying the reference safe models:* NAF assumes a safe model is given by an external entity. In a fair use case, if NAF is to be used, the defendant will need to produce a language model that (1) did not access the original copyrighted work and (2) outputs text with distribution close to the allegedly infringing language model. The fact that the defendant can choose different safe models for each prompt x and each original work c makes the notion of NAF unreliable. One with more resources could come up with better, safer models and claim a smaller NAF. This brittleness in the definition safe model leaves NAF open to criticism. The specific choice of the safe models and their implication in law, however, is beyond the scope of this work.

Our contributions. In this work, we aim to address the first and second challenges by proposing improved solutions and conducting comprehensive experiments to evaluate them empirically. We start by introducing the computation and estimation of NAF guarantees. To overcome the computation cost that scales linearly with the size of the output space, we propose a Monte Carlo estimator with a confidence bound that allows us to estimate NAF bounds for sentence-level generation.

Then, we compare the performance of CP-κ with a DP-based solution (trained with DP-FedAvg (McMahan et al., 2016)) on a next-token prediction task with an LSTM model[2] (Reddi et al., 2020). Next, we evaluate how CP-κ and CP-Δ can mitigate memorization in a fine-tuning task with a GPT-2 model (Radford et al., 2019). Finally, to further achieve a stricter NAF guarantee, we propose adding additional randomization into the generation process, such as increasing the decoding temperature, performing randomized response, or interpolating with an ε-DP model.

[2]Admittedly, the model is less expressive than modern transformer-based architectures given the constraints of resources for DP/federated training. However, this experiment effectively allows us to assess the trade-offs between NAF and utility.

References

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

Niva Elkin-Koren, Uri Hacohen, Roi Livni, and Shay Moran. Can copyright be reduced to privacy? *arXiv preprint arXiv:2305.14822*, 2023.

H Brendan McMahan, Eider Moore, Daniel Ramage, S Hampson, and BA Arcas. Communication-efficient learning of deep networks from decentralized data (2016). *arXiv preprint arXiv:1602.05629*, 2016.

U.S. Copyright Office. Copyright law of the united states (title 17), 2022. URL https://www.copyright.gov/title17/92chap1.html#107.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and H Brendan McMahan. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*, 2020.

Matthew Sag. The new legal landscape for text mining and machine learning. *J. Copyright Soc'y USA*, 66:291, 2018.

Sarah Scheffler, Eran Tromer, and Mayank Varia. Formalizing human ingenuity: A quantitative framework for copyright law's substantial similarity. In *Proceedings of the 2022 Symposium on Computer Science and Law*, pages 37–49, 2022.

Benjamin LW Sobel. Artificial intelligence's fair use crisis. *Colum. JL & Arts*, 41:45, 2017.

Nikhil Vyas, Sham M. Kakade, and Boaz Barak. On provable copyright protection for generative models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 35277–35299. PMLR, 23–29 Jul 2023.

Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. Detoxifying language models risks marginalizing minority voices. *arXiv preprint arXiv:2104.06390*, 2021.