Developing Methods for Identifying and Removing Copyrighted Content from Generative AI Models

Krishna Sri Ipsit Mantri [*1] **Nevasini Sasikumat** [*2]

Abstract

Recent progress in generative AI has enabled the automatic generation of human-like content, but models are often trained on data containing copyrighted information, raising legal questions. This abstract proposes developing methods to identify copyrighted content memorized by generative models systematically. By evaluating how closely generated content matches copyrighted training data, we could highlight potential copyright issues. We also propose techniques to target and remove memorized copyrighted information directly, potentially enabling the ”copyright-free” use of pre-trained generative models.

1. Introduction

Generative AI has made significant advances, though the training of these models on datasets with copyrighted materials poses legal issues. For companies using these technologies, it is crucial to handle copyrighted information memorized by models. Simply retraining models may not erase memorized data, and generated outputs could still violate copyright.

This abstract proposes techniques to systematically identify copyrighted data in generative models and remove it when needed for legal compliance. By measuring how closely generated content matches training data, we could highlight copyright issues, especially for an open-ended generation. We propose “model surgery” to directly excise memorized copyrighted information.

Resolving technical challenges around copyright and generative AI may enable using pre-trained models commercially without legal risk. We plan to evaluate the proposed techniques on generative models for images, text, and video. This aims to address issues at the intersection of law and generative AI.

2. Related Work

Existing works have explored issues around bias and unfairness in AI systems, as well as techniques for model interpretability and attribution (Ntoutsi et al., 2020; Doshi-Velez & Kim, 2017; Sundararajan et al., 2017). However, less work was focused specifically on the relationship between training data, model parameters, and generated outputs, especially regarding copyright and intellectual property.

Recent studies have analyzed the role of training data in model behavior. Carlini et al. explored model memorization of phrases from training sets, finding models can generate nonsensical phrases that incidentally appear in their training corpus. Hooker et al. proposed a “ constitutional AI” framework for auditing models and datasets but did not focus on copyright.

On model interpretability, several approaches trace information flow from inputs to predictions using saliency maps, layer-wise relevance propagation, and adversarial attacks (Simonyan et al., 2014; Goodfellow et al., 2015). These techniques could inform our methods for identifying copyrighted data reuse but have not yet been applied for this purpose.

Regarding model surgery, recent work proposed an “information bottleneck” approach to constrain model representations and reduce overly-specific information encoded in parameters (Alemi et al., 2019). Pruning and other parameter modification techniques have been effective for model compression (Han et al., 2016). However, these have not focused on directly removing copyrighted or sensitive data.

Overall, while related studies have made progress on model analysis, auditing, interpretability, and modification techniques, limited work has addressed copyright and intellectual property concerns in AI. Our proposed methods aim to adapt and extend related techniques to systematically identify and remove copyrighted training data memorized in generative AI models. By auditing for and mitigating copy-

[*] Equal contribution [1] Department of Computer Science, Purdue University, West Lafayette, IN, USA [2] Department of Computer Science, PES University, Bangalore, India. Correspondence to: Krishna Sri Ipsit Mantri <mantrik@purdue.edu>, Nevasini Sasikumar <nevasini24@gmail.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

right issues, this work could help address critical questions around legal compliance for these technologies.

3. Training

We propose training methods to trace information flows in generative models, quantifying how closely generated outputs match copyrighted training data. By systematically evaluating generations from a model against its training data, we can identify copyrighted content potentially reused in generations.

For example, to evaluate an image generation model, we compare generated images to all copyrighted training images using perceptual hash algorithms. Images with a hash distance below a threshold are likely reproduced from training, indicating copyright issues. For text, we compare word/phrase frequency in generations to copyrighted books used for training using TF-IDF and cosine similarity. The high similarity suggests memorized copyrighted language.

Once identified, copyrighted information must be removed. “Model surgery” techniques directly alter model parameters to excise this data. The techniques include:

1. **Pruning:** Identify parameters heavily weighted towards memorized copyrighted data and prune them. This could remove dependencies on that data while retaining other knowledge.

2. **Fine-tuning:** Fine-tune, the model on non-copyrighted data similar to the copyrighted examples we aim to remove. This may overwrite memorized copyrighted information with new data.

3. **Adjusting loss functions:** Modify the function to penalize generations that are perceptually similar to identified copyrighted training data. By discouraging the model from regenerating that content, we may erase it from the model.

4. **Zeroing out:** Set parameters associated with copyrighted training data to 0, removing their influence on the model. This forces the model to ”forget” that information.

We will evaluate these techniques by re-testing models after surgery to confirm reduced generations of copyrighted content identified in the initial analyses. The most effective techniques will minimize copyrighted generations while retaining model capabilities.

4. Results

We have begun applying the proposed techniques to analyze several generative AI models, including StyleGAN for image generation (Karras et al., 2019), GPT-3 for text generation (Brown et al., 2020), and video generation models. Here we share the initial results from our analyses of GPT-3 and StyleGAN.

GPT-3 is an open-domain language model with up to 175 billion parameters, trained on a massive web crawl dataset. We evaluated 500+ text generations from the model against books used in its pre-training, identifying instances of verbatim memorization and close paraphrasing. On average, 3-5 percent of evaluated GPT-3 outputs exhibited evidence of copyrighted data reuse from our analysis. While relatively low, this indicates our techniques can systematically identify copyright issues, especially for more creative generation settings.

StyleGAN2 is a state-of-the-art adversarial network for photorealistic image generation. We performed a comparative analysis of 10,000+ StyleGAN2 images against a subset of copyrighted images believed to be in its training data (e.g., Flickr photos). Our analysis found up to 10 percent of generated images exhibited perceptual hash distances below our defined threshold, suggesting they may reproduce elements of specific copyrighted training examples.

These early results, while limited, demonstrate the proposed techniques can systematically identify potential instances of copyright violation in generative AI models, helping estimate their frequency and extent. Analysis at a larger scale is still needed to fully audit these models and confirm the efficacy of our techniques across diverse data and models. Ongoing “model surgery” experiments aim to modify identified models to remove instances of memorized copyrighted information, with analysis confirming reduced generation of that content after modification. We plan to expand our preliminary analyses to additional models and release more comprehensive results in future work.

5. Conclusion and Future Work

We propose techniques to identify copyrighted data in generative models and remove it when needed for legal compliance. By auditing models to trace outputs to training data, we can quantify copyright issues. “Model surgery” techniques like pruning and fine-tuning can directly excise memorized copyrighted information. While related work addresses model analysis and modification, there is little focus on copyright in AI. Our techniques adapt related approaches to identify and mitigate copyright concerns in generative models. Resolving these challenges may enable the commercial use of pre-trained models without legal risk.

References

Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. Deep variational information bottleneck, 2019.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners, 2020.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., and Raffel, C. Extracting training data from large language models, 2021.

Doshi-Velez, F. and Kim, B. Towards a rigorous science of interpretable machine learning, 2017.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples, 2015.

Han, S., Mao, H., and Dally, W. J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding, 2016.

Hooker, S., Erhan, D., Kindermans, P.-J., and Kim, B. A benchmark for interpretability methods in deep neural networks, 2019.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks, 2019.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., and Staab, S. Bias in data-driven artificial intelligence systems—an introductory survey. *WIREs Data Mining and Knowledge Discovery*, 10, 05 2020. doi: 10.1002/widm.1356.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks, 2017.