Semantics-based Watermarking

Professor: Kannan Ramchandran (kannanr@berkeley.edu)
Students: Justin Kang, Syomantak Chaudhuri, John Wang, Nived Rajaraman
{syomantak, justin_kang, john-w, nived.rajaraman}@berkeley.edu

1 Abstract

The recent Executive Order on safe and trustworthy AI put out by the White Houses [7] lists watermarking as one of the guidelines for detecting AI-generated content and authentication of human-generated content. In the coming years, regulations for the use of detection and watermarking as a means to clearly label AI-generated content are likely to become stronger. Through this proposal, we aim to design new approaches for watermarking AI-generated content across different modalities (text/image/audio) which are robust to the natural and adversarial corruptions appearing in these domains.

2 Introduction

Generative models such as Large Language Models (LLMs) and diffusion models have become a powerful tool for automating systems and enabling multimodal content generation. However, the ever-increasing use of such models presents a number of concerns. Distinguishing between human-generated and AI-generated content has become more challenging than ever. Watermarking the output of the model is a promising approach toward resolving these issues: the model embeds a hidden signal (the watermark) in the output which can be identified by a detection algorithm. Watermarking techniques rely heavily on cryptographic primitives and one-way-functions to inject the hidden signal in a manner where it cannot be removed without changing the output of the model significantly. However, there is still a large gap toward designing watermarking schemes which can detect content generated from LLMs in an efficient, robust and reliable manner.

Classical approaches guarantee that the watermark can continue to be detected under a bound on how much the text/image/audio can be distorted by the user. However, in the modern era these constraints are less meaningful. Attacks such as the emoji attack [6] append to the prompt a request to the model to insert an emoji between every pair of tokens/words. The user then modifies the output of the model by removing the inserted emojis. In the process the watermarked text (with inserted emojis) is distorted significantly when all the inserted emojis are removed. It is thus clear that with current attacks, distortion constraints of the scale considered in classical watermarking approaches are violated many times over. Thus there is a necessity for new formulations and approaches for watermarking.

3 Technical objective

Natural corruptions of a model's outputs are more commonly present in image/video domains in various forms - subtle modifications such as compression/format conversion and more drastic ones such as cropping and filters on the image. On the other hand, adversarial corruptions by a user are often to conceal the fact that the text/image/audio was AI-generated and attempt to erase the watermark. Many of the current hashing-based watermarking approaches in the text modality hash the previous or previous-k tokens to generate the random seed for the current token being sampled. While these approaches can handle some amount of corruption, in the presence of attacks like the emoji attack, where the number of edits made to the output of the model is very large (approximately half the number of tokens generated), a large fraction of the hashes are thus recovered incorrectly. In addition to the emoji attack, another powerful attack against watermarking algorithms is the rephrasing attack, where the bad actor perturbs the model output by using a mask filling model to replace words/phrases with synonyms or semantically-equivalent expressions.

Proposed approach: The kind of perturbations appearing in natural and adversarial corruptions tend to be structured and not arbitrary. In particular, many existing attacks cause the model to output a simple transformation (insertions or edits) of the desired output. A natural question to ask is whether it is possible to design hash functions and thereby watermarking algorithms which are robust to these transformations. This is possible by designing robust and semantics-dependent hash functions which generalize in different modalities (namely text, image, and audio). Hash functions which depend on the semantics of the content

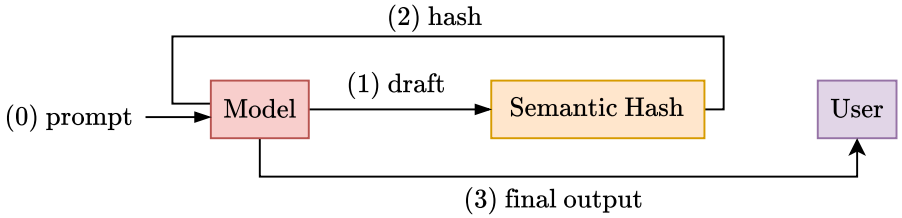

Figure 1: Semantics-based Watermarking

being processed are likely to be more robust to rephrasing attacks: although the sentence structure changes and words may be recast with their synonyms, but the meaning of the sentence largely stays the same. In particular, we consider designing robust hashing functions based on using modality-specific embedding models which map text, images, or audio into a fixed dimensional embedding. The embedding space aims to capture semantic similarity, with semantically similar generations closer to each other in cosine similarity. The high level approach we consider is detailed in Figure 1, where the idea is to use the semantics of a draft returned by the model to inject the watermark into the final output of the model. The robustness of the proposed watermarking scheme relies on validating the robustness of the semantic hash to natural and adversarial corruptions. Some questions to be addressed along the way to understanding the performance of this watermark,

- **Adversarial attacks:** In the text modality, how robust is the semantic hash to corruptions induced by rephrasing attacks and prompt injection attacks such as the emoji attack. Does the semantic hash associate a high cosine-similarity between the corrupted and uncorrupted text?
- **Natural attacks:** In the image/audio modality, how robust is semantic hashing to transformations such as compression and cropping/filtering?
- The proposed approach in Figure 1 consists of two forward passes through the generative model. Can smaller models be used for drafting? What are potential new adversarial attacks that can be leveraged against this watermark.

4 Real-world Applications of watermarking

Avoiding Training on Generated Inputs. LLMs are pre-trained on massive datasets. Such large uncurated datasets are likely to contain data generated from other LLMs. The effects of this self-consumption was studied in [1] where it was observed that too much machine generated content in the dataset negative effects model performance. Efficient watermarking allows this generated data to be excised from large pre-training datasets and reduce the extent of such systematic biases arising in the dataset. Furthermore, with more powerful techniques, other metadata can be injected as well, allowing for more fine-tuned curation.

Verifying Veracity. In certain setting, particularly those of video and images, there is always the concern that generated content may be passed-off as real-world content, resulting in the spread of misinformation. This difficulty has been highlighted in the recent White House executive order, and watermarking provides a potential solution to this problem.

Ensuring Accurate Attribution. We might also be interested in understanding to what extent a generative model was used to generate some data. For text specifically, the output of an LLM can be edited. When the volume of such edits grows, it falls into a grey area whether such text should be treated as being truly AI-generated or not. This presents an interesting problem for the next generation of watermaking solutions, which will need to go beyond a simple yes-no answer.

5 Related Work

Traditional approaches for watermarking embed a carefully crafted hidden signal in the content of digital data (such as images and software) [3, 4] without degrading the data itself and such that the detection is robust to some amount of corruption. However, watermarking for generative AI is different in one important way: the detection algorithm has access to the generative process of the text, image or audio being watermarked. Modern approaches such as Kuditipudi et al. [9] exploit this fact to watermark LLM-generated text in a way which does not degrade the quality of the text.

On the other hand, there has been a line of research developing watermarks which can be more robust at the cost of degrading the text quality. However, rather than distorting the text/image/audio directly, these approaches distort the distribution from which the image is sampled. In particular, the approach of Kirchenbauer et al. [8] distorts the distribution by increasing probability of a particular subset of the tokens (labelled as the "green" tokens) to be sampled based on previously generated text based on a known hash function. For detection of watermark, the algorithm checks if there are significantly more green tokens than expected from non-watermarked text. There are a number of other recent works which introduce new watermarking approaches across different modalities [10, 11, 2, 12, 5].

References

[1] Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G Baraniuk. Self-consuming generative models go mad. *arXiv preprint arXiv:2307.01850*, 2023.

[2] Bang An, Mucong Ding, Tahseen Rabbani, Aakriti Agrawal, Yuancheng Xu, Chenghao Deng, Sicheng Zhu, Abdirisak Mohamed, Yuxin Wen, Tom Goldstein, and Furong Huang. Benchmarking the robustness of image watermarks, 2024.

[3] Dan Boneh and James Shaw. Collusion-secure fingerprinting for digital data. *IEEE Transactions on Information Theory*, 44(5):1897–1905, 1998.

[4] Jim Chou, S Sandeep Pradhan, Laurent El Ghaoui, and Kannan Ramchandran. Watermarking based on duality with distributed source coding and robust optimization principles. In *Proceedings 2000 International Conference on Image Processing (Cat. No. 00CH37101)*, volume 1, pages 585–588. IEEE, 2000.

[5] Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models, 2023.

[6] Riley Goodside. There are adversarial attacks for that proposal as well — in particular, generating with emojis after words and then removing them before submitting defeats it. 2023.

[7] The White House. Executive order on safe, secure, and trustworthy artificial intelligence. 2023.

[8] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International Conference on Machine Learning*, pages 17061–17084. PMLR, 2023.

[9] Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models. *arXiv preprint arXiv:2307.15593*, 2023.

[10] Robin San Roman, Pierre Fernandez, Alexandre Défossez, Teddy Furon, Tuan Tran, and Hady Elsahar. Proactive detection of voice cloning with localized watermarking. *arXiv preprint arXiv:2401.17264*, 2024.

[11] Pushmeet Kohli Sven Gowal. Synthid: Identifying ai-generated content with synthid. 2023.

[12] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust, 2023.