Data Mining (CS F415) : Semester 2, AY 2023-24 Group 7, Project SNo 33
Suraj Phalod, Chaitanya Sethi, Dhawal Mehta https://arxiv.org/abs/2305.20030

TREE-RING WATERMARKS: FINGERPRINTS FOR DIFFUSION IMAGES THAT ARE INVISIBLE AND ROBUST

YUXIN WEN, JOHN KIRCHENBAUER, JONAS GEIPING, TOM GOLDSTEIN

Abstract. The authors have proposed a watermarking technique for generative models, specifically for diffusion models. They call it Tree-Ring Watermarking, a novel technique for robustly fingerprinting diffusion model outputs. Unlike other models which propose watermarking as a post processing step, this paper proposes entering the latent space, watermarking it at that initial noise vector and then letting the diffusion model progress, thus rendering the watermark imperceptible to human eyes. These patterns are carefully designed in Fourier space, ensuring their invariance to common image transformations such as convolutions, crops, dilations, flips, and rotations. After image generation, the watermark signal is extracted by reversing the diffusion process, allowing retrieval of the noise vector. Subsequently, the noise vector is examined for the presence of the embedded signal. This technique can be easily applied to arbitrary diffusion models, including text-conditioned Stable Diffusion. Watermarking the outputs of generative models is a crucial technique for tracing copyright and preventing potential harm from AI-generated content.

1. Understanding of the Paper

Fundamentally, the process of Tree-Ring watermarking is as follows

(1) The model has pre-trained weights unknown to the user.
(2) User calls the model through the Tree-Ring Watermark API.
(3) The model samples an initial noise vector.
(4) The model watermarks the noise vector after performing the FFT. This is done using some kind of a key.
(5) Then, they use the model to produce an image with the perturbed noise vector.
(6) This image is returned back to the user.
(7) In case the user uses it for malicious intent, or to verify and traceback the original model and whether it was used to create the image, we go back to the model.
(8) We use the reverse diffusion process to find the noise vector from which the image was created.
(9) If we find the watermark key in the noise vector, Fourier transformed space, then we know that it has been created using this particular model.
(10) This, obviously, holds under the assumption that no one else knows the weights of the training model.

The next important aspect to understand is, HOW do we watermark?

(1) We produce an initial noise, in this case an initial Gaussian sample
(2) Run a Fast Fourier Transform to Fourier space. (NOTE: Gaussians in Fourier space are just Gaussians)
(3) Now, we add the pre-defined keys in the latent, FFT space. Their pre-defined keys are of various types, as described in the paper (Zeros, Rand, Rings). They finally went ahead with a multi-ring, concentric key, following a pattern. Imprint that as rings in the FFT space, as -
    (a) A rotation in pixel space corresponds to a rotation in Fourier space.
    (b) A translation in pixel space multiplies all Fourier coefficients by a constant complex number.
    (c) A dilation/compression in pixel space corresponds to a compression/dilation in Fourier space.
    (d) Color jitter in pixels (adding a constant to all pixels in a channel) corresponds to changing the magnitude of zero-frequency Fourier mode.
(4) We then do Inverse Fourier Transform, after adding the key to FFT space.
(5) The noise that we get back is no longer pure Gaussian noise of the original intended distribution, however it proves to be a good enough distribution.
(6) We then run the normal DDIM diffusion model, and we get the output. There is no post-processing required, this image is already watermarked.

OPTIONAL We can perturb the image with a range of operations, such as compression, blurring, rotation, etc. and the watermark holds in Fourier space.

(7) We then perform the inverse process, using DDIM inversion, and again apply FFT to reach inverted Fourier Space (model weights required)
(8) We then read out the key. If the image was produced by our target model, then on running the prediction algorithm, we should be able to identify it with the key we implanted. This can be used to identify the user/key as well as the model/API used for the same.

2. Conclusion and Future Work according to our understanding

Conclusion

Tree-Ring watermarking is a novel approach to watermarking for diffusion images based on patterns in the Fourier space of the noise vector. It uses minimal but optimal shifts of the output distribution making it invisible on a per sample basis, and making it robust to strong perturbations and transformations.Tree-ring watermark remains detectable even under strong image manipulations that might be faced in malicious usage and handling of generated images.

Future Work

Currently we haven't trained the diffusion model on the modified latent space after adding the key in the Fourier transformed space. Can we train the diffusion model that has watermarked noises as an actual training distribution? Even after watermarking, there might be a an approach to steer the outputs, which has not been tried yet. Increasing the accuracy of the model, and finding new ways to watermark in such a way that it is invisible to the human eye and resistant to perturbations.