Preliminary Research Idea (13): Using Diffusion Repainting Mechanism to Achieve Cross-Domain Thinking in Image and Text Watermarking

Professor Richard Yi Da Xu
HKBU and TadReamk Limited

Introduction

Recently, I read the TRELLIS+ paper, which integrates multiple technologies including Dino, Gaussian Splatting+, Radiance Field, Flow Matching+, and CLIP+. I found great satisfaction in understanding how these concepts are integrated. This has inspired me to think about applying Diffusion Repainting+ mechanism to watermarking tasks.

Core Research Idea

The main hypothesis is: Can we use the Repainting mechanism during the diffusion process (at different time steps t with a carefully designed Mask Pattern) to achieve image watermarking? Is this idea novel, or has it already been explored? Furthermore, can this approach be migrated to Discrete Diffusion Models in Natural Language Processing (NLP) for text watermarking?

Technical Explanation of Diffusion Repainting

Diffusion Repainting generates image x_0 from pure noise x_T. In image restoration, an image is divided into:

- **Known region (Unmasked):** Pixels to be preserved (Ground Truth).

- **Unknown region (Masked):** Missing pixels to be generated and blended.

Step A: Model Prediction (for Unknown Regions)

During denoising from time step t to $t-1$ (a cleaner state), the diffusion model observes the current noisy state x_t to predict x_{t-1}. The model generates content for the entire image, including known parts. This prediction is called x_{pred}.

Note that directly sampling from x_T without known region constraints would result in a new image, which is undesirable. This implies the need to integrate back the known regions.

Step B: Hard Constraints (Targeting Known Regions)

1. Take out your original, clean real image (x_{clean}).

2. Use forward diffusion to add noise at each step t (of course, including matching the noise level of step $t-1$). We call this $x_{\text{known noisy}}$.

3. Combine them:

   - For masked (unknown) regions: Retain pixels from the model's prediction (x_{pred}).
   - For unmasked (known) regions: Discard the model's prediction and paste in $x_{\text{known noisy}}$.

Step C: "Redrawing"

If you only perform Step A and B once at each step, the boundary between "pasted known pixels" and "generated unknown pixels" might appear discontinuous or full of seams. This is because the model doesn't have enough time to coordinate them. So we solve this problem by redrawing through cyclic time:

1. **Denoising:** From t to $t-1$ (using the combination method above).

2. **Re-add noise (re-bounce):** Add noise to the entire image back to step t.

3. **Denoise again:** Denoise the entire image again from t to $t-1$.

Through this "two steps forward, one step back" dance, the model gets multiple opportunities to observe the "known" pixels you pasted in, and adjust the "unknown" pixels to perfectly match the boundaries.

Key point: Step C can be used to remove Boundary Artifacts. This idea feels quite general and might be applicable to other image artifact processing areas.

Application to Watermarking

The proposed approach involves using the Repainting mechanism with carefully designed mask patterns at different diffusion time steps to embed watermarks into images. The same principle could potentially be extended to discrete diffusion models in NLP for text watermarking, where certain tokens or regions could be treated as "known" (watermark) and others as "unknown" (to be generated while preserving the watermark).