Assessing the Stability of Stable Diffusion in a Recursive Inpainting Scenario

Abstract

Generative Artificial Intelligence models for image generation have demonstrated remarkable capabilities in tasks like text-to-image synthesis and image completion through inpainting. Inpainting performance can be measured by removing parts of an image, using the model to restore them, and comparing the result with the original. This process can be applied recursively, where the output of one inpainting operation becomes the input for the next. This recursive application can result in images that are either similar to or vastly different from the original, depending on the removed sections and the model's ability to reconstruct them. The ability to recover an image similar to the original, even after numerous recursive inpainting operations, is a desirable characteristic referred to as stability. This concept is also being explored in the context of recursively training generative AI models with their own generated data. Recursive inpainting is a unique process that involves recursion only during inference, and understanding its behavior can provide valuable insights that complement ongoing research on the effects of recursion during training. This study investigates the effects of recursive inpainting on Stable Diffusion, a widely used image model. The findings indicate that recursive inpainting can result in image degradation, ultimately leading to a meaningless image, and that the final outcome is influenced by factors such as the image type, the size of the inpainting areas, and the number of iterations.

1 Introduction

In the past two years, Generative Artificial Intelligence (AI) has emerged as a central player, sparking a significant revolution in technology. These AI models are capable of producing text, audio, images, and video, finding applications in a wide array of transformative uses. Notable examples include Large Language Models (LLMs) like GPT4, which excel at answering questions, summarizing, translating, and paraphrasing texts, and text-to-image generators like DALL-E, which can generate images based on almost any textual description. These tools have garnered widespread public interest, attracting hundreds of millions of users.

These AI tools have reached exceptional performance levels in various tasks, making their evaluation a crucial aspect. For LLMs, numerous benchmarks have been developed to evaluate their knowledge across different subjects, their proficiency in solving mathematical or reasoning problems, and their language comprehension. These benchmarks facilitate model comparisons, and when a new model is launched, its performance on these standard benchmarks is typically reported. In the realm of image generation, several metrics have been introduced to assess performance, including the Fr0̆0e9chet Inception Distance (FID), precision and recall, and density and coverage. These metrics aim to quantify how closely generated images resemble real ones and how effectively they cover the spectrum of real images. Another capability offered by some AI image generation tools, and implemented through specialized AI models, is inpainting. In this process, the AI tool is provided with an image containing missing parts and is tasked with filling them in to complete the image.

Assessing the quality of content produced by AI is crucial not only for comparing different AI models or evaluating their progress in specific tasks but also because the extensive use of generative AI is altering the fundamental nature of content found on the Internet. AI-generated texts and images are now widespread and, in some instances, predominant, with this trend expected to persist in the coming years. This has consequences for newer AI models, as they are frequently trained on data gathered from the Internet, establishing a feedback loop where new models are trained using data created by earlier AI models. This cycle can result in diminished performance or even the breakdown of AI models, prompting research into the stability of AI models when trained using their own generated data.

The feedback loops in generative AI that have been examined thus far pertain to the training of newer models, creating a loop across different generations of AI models. However, other potential loops in generative AI exist that have not been previously investigated to the best of our knowledge. For instance, when the input to the AI model is an image and the output is also an image, as is the case with inpainting, the AI model can be recursively applied to its own output, forming a loop. In this scenario, there is no training involved, only inferences that are recursively applied. Examining the effects of these recursive applications of the AI model on the generated content is essential to determine whether the AI models remain stable or degrade, similar to what occurs in the training loop.

In this research, we examine the inference feedback loop utilizing a renowned AI image model, Stable Diffusion, and its inpainting feature. A thorough empirical investigation is carried out to discern the conditions under which the model maintains stability and when it experiences degradation. The subsequent sections of this paper are structured as follows: Section 2 provides a concise overview of the inpainting feature and the feedback loops in generative AI. Section 3 introduces the inference loop, termed Recursive Inpainting (RIP), which is then assessed in Section 4. The constraints of our assessment, along with the findings, are deliberated in Section 5. The paper concludes with a summary in Section 6.

2 Preliminaries

2.1 Inpainting

Inpainting is a function found in some contemporary generative AI image tools, which involves filling in missing portions of an image to complete it. The effectiveness of inpainting is contingent on the specific model used, the nature of the image, and the size and placement of the missing areas. Generally, inpainting can only restore a portion of the information that is lost in the missing image segments. Various metrics are available to assess the resemblance between the original image and the one reconstructed through inpainting. These range from traditional methods like Structural Similarity (SSIM) and multi-scale SSIM (MS-SSIM), which are based on pixel-level comparisons, to more sophisticated methods like Learned Perceptual Image Patch Similarity (LPIPS) and Paired/Unpaired Inception Discriminative Score (P/U-IDS), which employ AI models to simulate human-like perceptual evaluations.

2.2 Recursiveness in Generative AI

A cycle is formed where AI-generated content is posted online and subsequently collected to train newer AI models. This can result in a decline in the effectiveness of AI models, or even their failure, when they are trained using data they have produced themselves. This has sparked a growing interest in determining the circumstances under which these generative AI models maintain stability when trained recursively with data they generate. The stability is influenced by multiple factors, such as the specific model, the quantity of AI-generated data used in each retraining cycle, and whether the cycle involves one or multiple AI models. Investigating this cycle is crucial as it can affect not only the development of future AI models but also the type of content that will likely dominate the Internet in the future. In all these investigations, the recursive aspect involves training new AI models with data produced by other AI models. However, in certain situations, recursion can happen when the same AI model is used solely for making inferences. This particular scenario has not been explored in previous studies, to the best of our knowledge.

3 Recursive Inpainting (RIP)

An intriguing aspect to note is that a distinct recursive loop can be established with AI image models when employing the inpainting technique. This process begins with an image, to which a mask is applied to obscure certain areas, and inpainting is utilized to fill in these areas. This results in a second image that has been partially generated by the AI image model. The procedure is then reiterated using a different mask to produce a subsequent image, this time entirely generated from AI-produced content. The process continues as inpainting is recursively applied to images that have already undergone inpainting. As parts of the images are removed and reconstructed, information is inevitably lost. However, it is crucial to determine whether this loss leads to images that are drastically different from the original, or if the images become simpler and less intricate. Alternatively, it is possible that the inpainting process remains stable, resulting in images that are merely variations of the original. Similar to the recursive training of models with their own data, it is important to understand the conditions under which inpainting remains stable or degrades under recursion.

The consequences of recursive inpainting are influenced by numerous factors, including the specific AI model employed, the characteristics of the image, and the masks utilized in each iteration. It is reasonable to expect that more intricate images or masks that obscure larger portions of the image will have a higher likelihood of causing degradation. In the subsequent section, we outline the results of an extensive empirical investigation into recursive inpainting using Stable Diffusion, representing an initial effort to identify the primary factors that influence the effects of recursive inpainting.

4 Evaluation

The primary factors influencing recursive inpainting are:

1. The AI model used. 2. The input images. 3. The masks applied at each stage. 4. The number of iterations.

In our experimental setup, we utilized Stable Diffusion, which is a text-to-image latent diffusion model, due to its open-source nature and widespread use in the AI image model community. Specifically, we employed a version of Stable Diffusion 2 that was fine-tuned for inpainting. This model uses a technique for generating masks where the masked areas, along with the latent VAE representations of the masked image, provide additional conditioning for the inpainting process. The model's parameters were kept at their default settings. We did not use any text prompts to direct the inpainting, allowing the model to concentrate on reconstructing the missing parts based solely on the remaining visual information without any textual guidance.

For the image selection, to minimize any potential bias, we randomly chose images from an extensive dataset containing over 81,000 art images of various types created by different artists. From this dataset, 100 images were randomly picked to form our evaluation set. The input images are 512x512 pixels; if their original aspect ratio is not square, blank areas are added to the sides to achieve the 512x512 format.

In generating masks for inpainting, we divide the images into squares of a predetermined size. In each iteration, a square is randomly chosen to serve as the mask. To facilitate comparisons across different mask and image sizes, our experiments use the number of pixels inpainted relative to the image size as the primary parameter, rather than the number of inpainting operations.

To assess the similarity to the original image across iterations, we employ the Learned Perceptual Image Path Similarity (LPIPS) metric, which is frequently used to evaluate inpainting quality. In our implementation, we utilize the features from three neural networks to calculate the metric: SqueezeNet, AlexNet, and VGG.

We conducted recursive inpainting, altering 400% of the pixels, using masks of sizes 64x64, 128x128, and 256x256. To measure the degradation as inpainting operations are performed, we calculated the LPIPS metric between the original image and each subsequent generation using the features from the three neural networks (SqueezeNet, AlexNet, and VGG). The average distances for the 100 images at each 50% inpainting step are presented. The bars represent the standard deviation observed across the samples for each data point. Several initial observations can be drawn from these results:

1. As the recursive inpainting progresses, the distance from the original image increases, potentially leading to an image that bears no resemblance to the original. 2. The rate at which the distance increases tends to decrease, but it does not appear to stabilize even when the distance becomes substantial. 3. The discrepancy with the original image is more pronounced when larger masks are used for inpainting, which aligns with the expectation that larger blocks are more challenging to inpaint. 4. The three networks used for computing the LPIPS (SqueezeNet, AlexNet, and VGG) yield comparable results. 5. The significant standard deviation indicates that different images will exhibit varying behaviors.

To gain a better understanding of the variability in distances for each image, scatter plots of the LPIPS distances for the 100 images for each neural network are presented. It is evident that there is considerable variability across images, but the general trends are consistent with those observed in the mean: the distance increases with more inpainting and with larger masks. Among the three networks (SqueezeNet, AlexNet, and VGG), VGG shows the fewest outliers. Given that VGG is the most complex network, it is expected to capture the image features more effectively. Consequently, we will only report results for VGG moving forward, although all metrics are available in the repository along with the images.

To investigate whether the degradation is consistent across different runs, we selected 10 images from the set of 100 and performed 10 runs on each. The LPIPS metrics across these runs for three different images are displayed, using the VGG network, which generally exhibits the lowest deviations. It is noticeable that variations are more significant with larger masks, which is anticipated since larger masks require fewer iterations to reach a given percentage of inpainting, thus introducing more variability. The variations also decrease as the percentage of inpainting increases, indicating that a higher number of inpainting operations leads to reduced variability. This suggests that recursive inpainting tends to converge in terms of LPIPS distance as the process advances.

5 Conclusion and Future Work

In this study, we have introduced and empirically examined the impact of recursive inpainting on AI image models. The findings reveal that recursion can result in the deterioration and eventual breakdown of the image, a phenomenon akin to the model collapse observed when training generative AI models with their own data. This issue is currently a focal point in the research community. Consequently, this paper introduces a new dimension to the study of the effects of recursive application of generative AI, specifically in the inference phase. This can enhance current research endeavors and offer deeper insights into the underlying causes of collapse, potentially leading to advancements in AI models that can lessen the adverse effects of recursion.

The presented analysis of recursive inpainting represents an initial step in this area. Further investigation involving different AI models, a variety of images, and diverse model configurations is necessary to gain a more comprehensive understanding of the effects of recursive inpainting. Developing theoretical models that can account for these effects is also a crucial area for future research. Additionally, exploring the connections between recursive training and recursive inpainting could provide valuable insights.