ANDRA ALĂZĂROAIE
A.ALAZAROAIE-1@STUDENT.TUDELFT.NL

Final project course: CSE3000 Research Project
Responsible Professor and Supervisor: Anna Lukina

The Many Faces of AI Art: Self-Poisoning Generative Models

01. Introduction

The Autophagous (Self-Consuming Training Loop)[1]: *AI models use their own generated data for training.*
- ***Degeneration***
- ***Loss of Diversity***
- ***Decreased quality***

Generative AI Models and Art: *Can AI produce creative art?*
Creative Novelty assessment method: *based on focal subject matter and its intrerrelations within the artwork [2].*

Content Novelty:
- *Image Description: Generate text descriptions using BLIP-2 [3].*
- *Text Vectorization: Convert descriptions into vectors with BERT [4].*

Visual Novelty:
- *Image Vectorization: Create image vectors using DINOv2 [5].*

Comparison: *cosine distance between the vector representations of the new and baseline artifacts.*

02. Contribution

*Defining and investigating **self-poisoning**.*

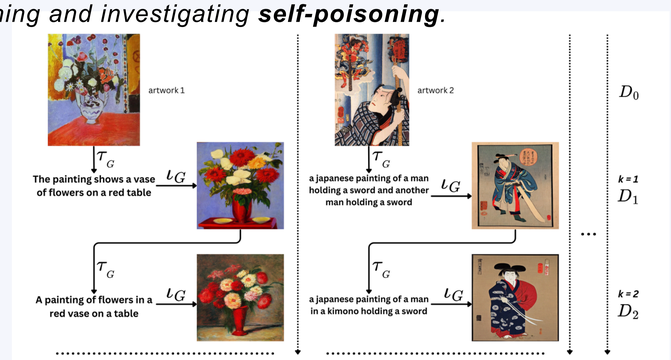

Figure 1: Process Flow Diagram illustrating the Self-Poisoning Process on an individual artwork level, containing particular examples.

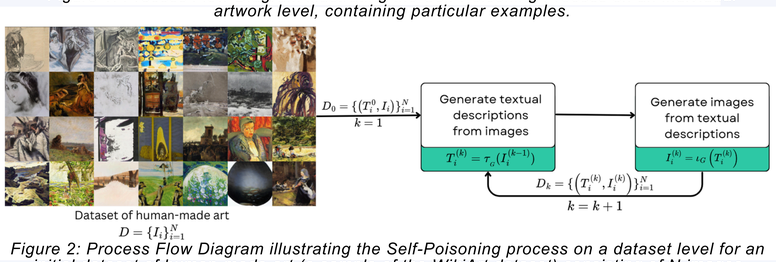

Figure 2: Process Flow Diagram illustrating the Self-Poisoning process on a dataset level for an initial dataset of human-made art (a sample of the WikiArt dataset) consisting of N images

Main Question: *How does the iterative process of generating images from text descriptions, and vice versa, affect the novelty and quality of the outputs?*

Sub-Questions:
- *Content Drift*
- *Visual Changes*
- *Convergence*
- *Impact on Quality*

Image-to-Text model: *BLIP-2 [3]*
Text-to-Image model: *Stable Diffusion [6]*
Dataset: *sample of 270 images from WikiArt [7] (10 random images from each artistic style)*

03. Results

Increase in Content and Visual Novelty *(relative to the human-made artworks).* ***Convergence*** *at 66th/67th iteration.*

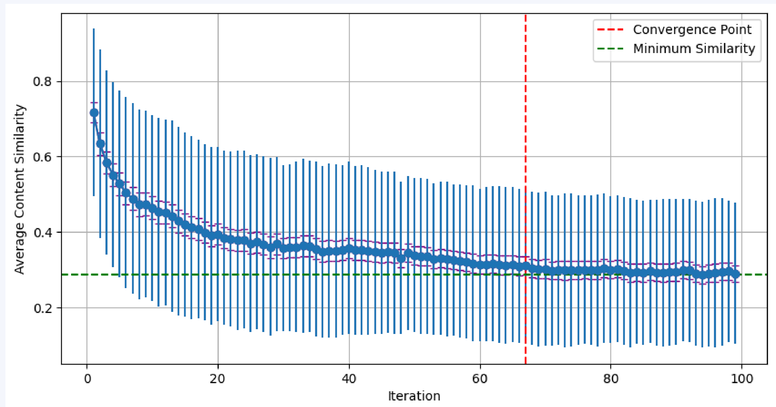

Figure 3: Content Similarity between the first and the other generated captions over iterations. The blue points represent the average Content Similarity, the blue error bars represent the standard deviation, and the purple error bars represent the 95% confidence interval.

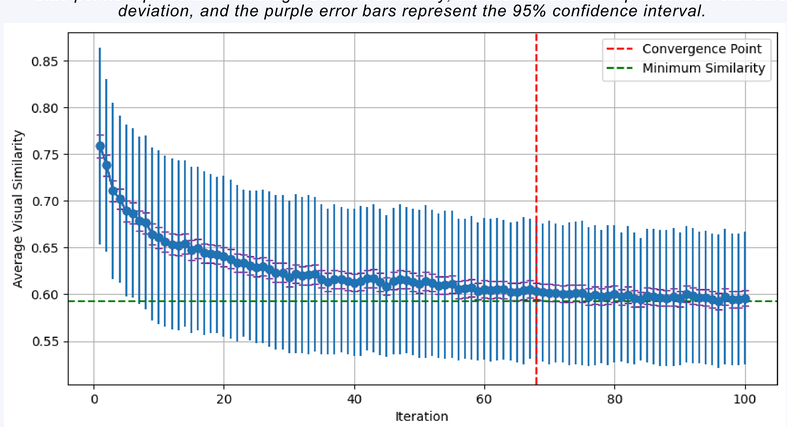

Figure 4: Visual Similarity between the initial artwork and the generated images over iterations. The blue points represent the average Content Similarity, the blue error bars represent the standard deviation, and the purple error bars represent the 95% confidence interval.

Decrease in Quality *(relative to the human-made artworks): measured through the Fréchet Inception Distance metric.*

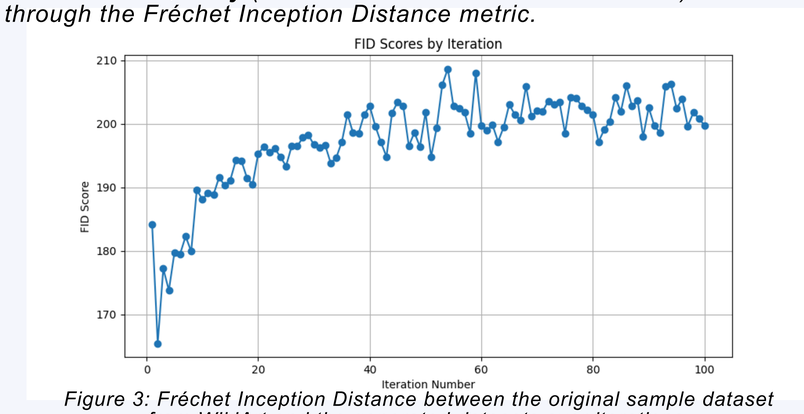

Figure 3: Fréchet Inception Distance between the original sample dataset from WikiArt and the generated datasets over iterations.

Qualitative Analysis:
- *an example with little novelty, but existing connection to the original artwork*
- *an example with a lot of novelty, but divergence from the intent of the artwork*

Figure 5: Self-Poisoning Example 1, starting from Adam Baltatu's artwork "House With Flowers" in style Post Impressionism.

Figure 6: Self-Poisoning Example 2, starting from Edward Hopper's ar.twork "Automat 1927" in style New Realism

04. Limitations

1. ***Dataset:*** *Only 270 images from WikiArt due to hardware and time constraints.*
2. ***Iterative Depth:*** *100 iterations conducted, more iterations for better understanding of long-term effects.*
3. ***Models:*** *Only used BLIP-2 and Stable Diffusion. Testing with more models could improve generalizability.*
4. ***Creativity Framework:*** *Based on a single prior study.*

05. Conclusions

Findings: *self-poisoning*
- Introduces content and visual novelty.
- Degrades quality over time.
- Convergences on certain themes and styles.

Implications:
- Stable Diffusion [6] and BLIP-2 [3] fail to accurately and completely transpose their input into their output.
- There are limitations in preserving the original complexity and creativity.
- Yet, novel elements arise inspiring from previous artworks.
- Enhancing models' capabilities of capturing and replicating complex artistic elements is needed for advancing AI art generation.

06. References

[1] S. Alemohammad, J. Casco-Rodriguez, L. Luzi, A. I. Humayun, H. Babaei, D. LeJeune, A. Siahkoohi, and R. G. Baraniuk, "Self-consuming generative models go mad," arXiv preprint arXiv:2307.01850, 2023.
[2] E. Zhou and D. Lee, "Generative artificial intelligence, human creativity, and art," PNAS Nexus, vol. 3, no. 3, p. pgae052, 03 2024. [Online]. Available: https://doi.org/10.1093/pnasnexus/pgae052
[3] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre training with frozen image encoders and large language models," in International conference on machine learning. PMLR, 2023, pp. 19 730–19 742.
[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.
[5] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby et al., "Dinov2: Learning robust visual features without supervision," arXiv preprint arXiv:2304.07193, 2023.
[6] CompVis, "Stable diffusion," https://github.com/CompVis/stable-diffusion, 2022, accessed: 2024-04-23.
[7] "Wikiart," https://www.wikiart.org/, accessed: 2024-06-03.