Do you remember? Memorization in diffusion models

02501 Advanced Deep Learning in Computer Vision

Aasa Feragen
Professor, DTU Compute

12 March, 2024

Keywords/lectures: Diffusion models, memorization

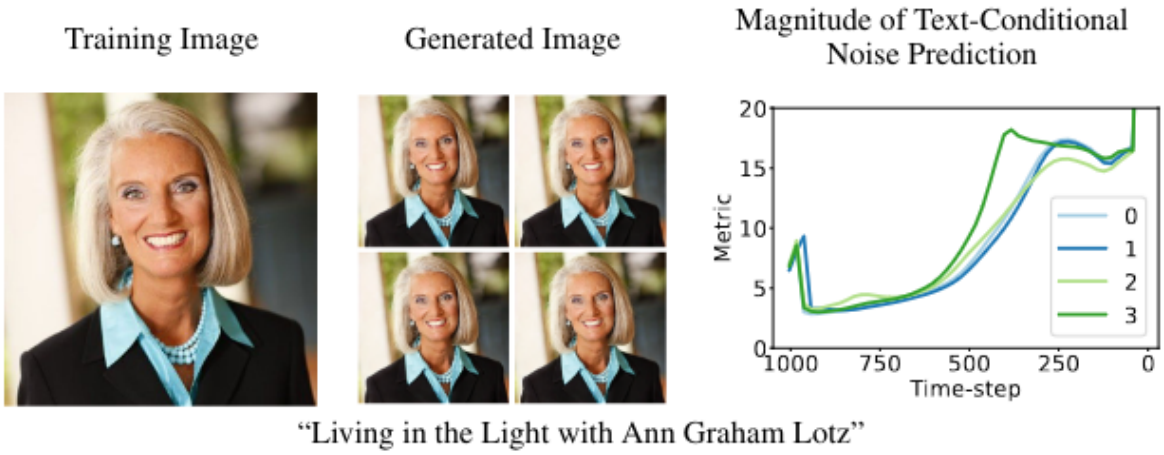

Figure 1: **Memorization in diffusion models.** Generative models have been known to memorize – diffusion models are no exception. This comes, however, at great risk for violating both property law and privacy when generative models are used in off the shelf software.

1 Project description

Memorization refers to the phenomenon of generative models learning to reproduce their training data, or elements of it. When generative models are used to generate text or images, this comes with a risk of users committing plagiarism unknowingly. When generative models are used in healthcare application, this comes with a risk of breaking privacy regulations. Hence, detecting and mitigating memorization is of crucial importance.

In this project, you will train diffusion models on medical imaging datasets and analyze them to identify and mitigate potential memorization.

2 Data/Resources

You will train an diffusion model with and without classifier-free guidance on the CheXpert chest X-ray dataset [1].

3 Tasks

In this project, you could work on the following tasks:

Task 1: Implement and train the diffusion model on CheXpert with and without classifier free guidance. Implement the diffusion model and train it with and without classifier free guidance to enable conditioning on a range of diseases. but still just one model?

Task 2: Implement and validate testing for memorization. Implement the method from [2] to detect signs of memorization. Validate its results both visually and using semantic metrics to assess whether the memorization metric correlates with semantic similarity to training data those with high score and similarity in train? plotting but also just 1 corr score?

Task 3: Mitigate memorization If you have detected memorization, implement some of the mitigation techniques from [2] and repeat your validation to assess their effect.

one is by modifying high trigger token:
1. inference-time mitigation: adjust embedding based on loss
2. training-time mitigation: remove high magnitude prompt from batch

References

[1] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI*, volume 33, pages 590–597, 2019.

[2] Yuxin Wen, Yuchen Liu, Chen Chen, and Lingjuan Lyu. Detecting, explaining, and mitigating memorization in diffusion models. In *ICLR*, 2023.