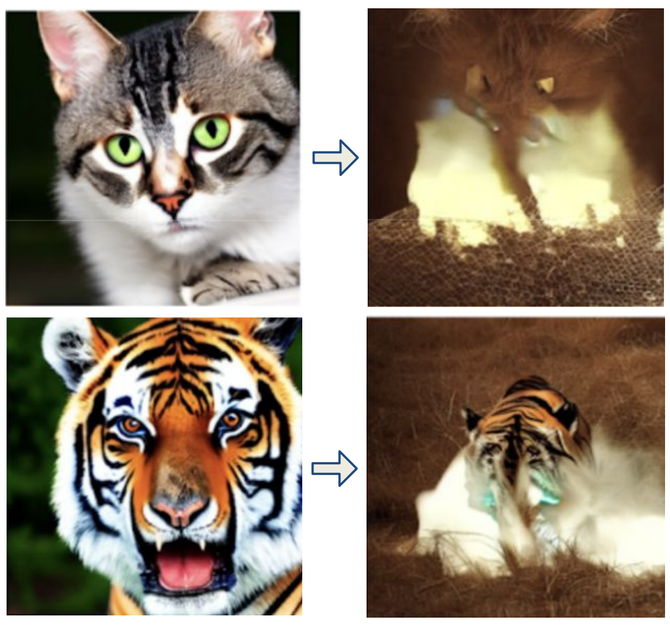
Evaluating Machine Unlearning Methods Using Sparse Autoencoders

Paul Ellement
University of Toronto

Motivation

- **What is machine unlearning?** Removing specific knowledge (e.g., cats, Van Gogh) from an AI model to prevent certain outputs
- **What are its applications?** Unlearning in text-to-image diffusion models is important to prevent generating images that contain:
  - Copyrighted material (e.g., a company's mascot)
  - Private data (e.g., a real person's face)
  - Inappropriate/violent images (e.g., nudity, self-harm)
- **Unlearning has two main problems:**
  1. Erasing neighbouring concepts (e.g., unlearning cats also leads to unlearning lions, leading to lower model utility)
  2. Concept is not completely erased

Images Before vs. After Unlearning "cat" [1]



Related Work

- **Sparse Autoencoders (SAEs):**
  - **What are they?** A model which maps the original model's activations to sparse latents, such that one concept will activate as few neurons as possible.
  - **How are they used in Unlearning?** Since concepts can be attributed to certain SAE neurons, these can be used to change weights of original model to avoid activating these neurons [1]
- **Evaluating Concept Unlearning**
  - EraseBench [2] is a benchmark for assessing unlearning, which provides hundreds of sets of test prompts, and metrics to assess both effectiveness and side effects of erasure.
- **Improving Erasure Methods**
  - Methods of erasure besides SAEs include fine-tuning and censored-training (retraining without explicit data) [3].

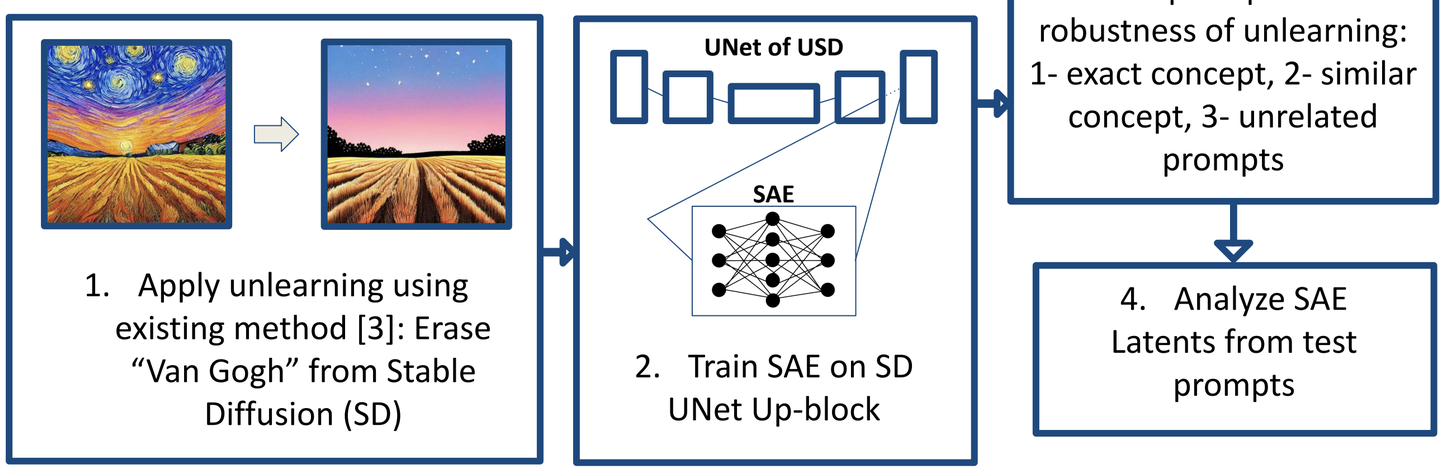
References

[1] Cywinski & Deja. Interpretable Concept Unlearning in Diffusion Models with SAEs. Research Gate, 2025.
[2] Amara et al. How Concept Erasure Degrades the Generation of Non-Target Concepts. Arxiv, 2025.
[3] Gandikota et al. Erasing Concepts from Diffusion Models. ICCV, 2023.

New Technique

- **Unique approach:** SAEs are useful for model interpretability, therefore they can be used for interpreting how effective a machine unlearning method is.
- **We want to find out:**
  - Is there residual knowledge of the concept after unlearning?
    - Can a unique **latent direction** (i.e., one or a few neurons) associated with the concept be identified?
  - What were the unintended consequences of the erasure?
    - How do the latents (the SAE activations) of prompts related to the concept compare to the latents of similar concepts?

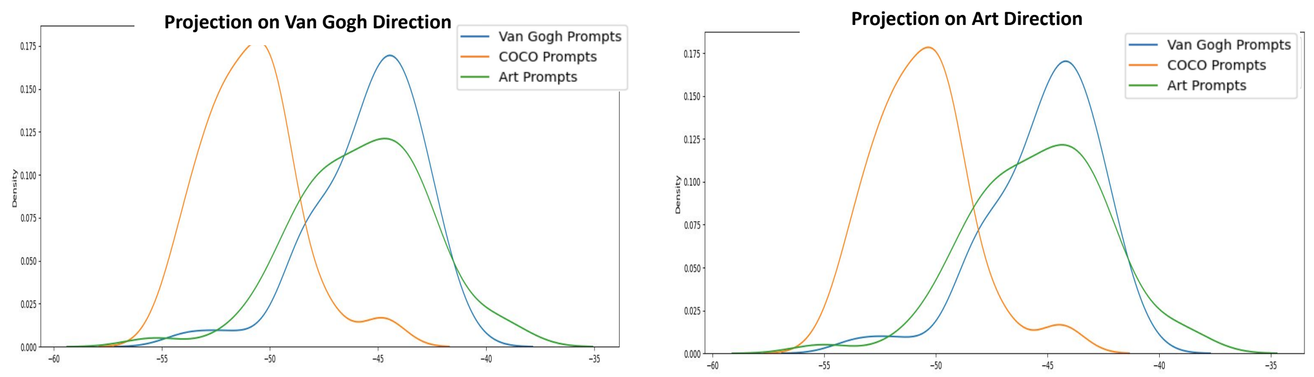
Overview of Methodology



Experimental Results

Overall Results: Since SAEs can't perfectly capture model's representations, 1- (similar to EraseBench [2]) erasure is not precise, so non-target concepts are impacted, and as a result, 2- it's not an ideal tool to use to conclude whether the unlearning was effective. However, statistical analysis shows signs of residual knowledge of the unlearned concept.

Projection of Prompts onto Various Directions



Projection of various types of prompts onto the Van Gogh Direction (left) and art direction (right). Similarity between both plots, and between Van Gogh and art prompt distributions indicate similar direction; indicating some erasure of concept

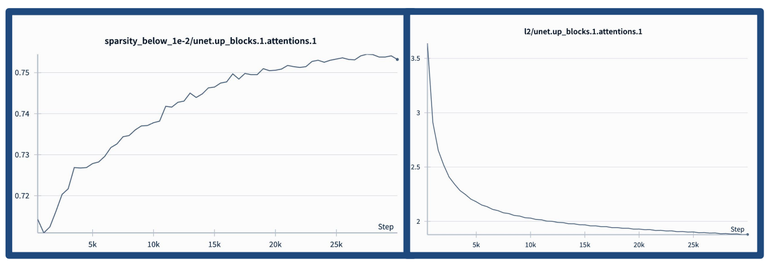
Base vs. SAE output images

Comparison of images generated with base unlearned model (left) to base model with SAE hookpoint (right) for prompt "Image of a cat" (top) and "Painting in the style of Van Gogh" (bottom)

Cosine Similarity Between Directions

Directions	Cosine Similarity
Van Gogh vs. Art	0.9947
Van Gogh vs. (Van Gogh – Art)	0.4698
Art vs. (Van Gogh – Art)	0.3768

SAE Training Metrics



Training metrics of SAE. Sparsity (left) should be high, reconstruction loss (right) should be low.