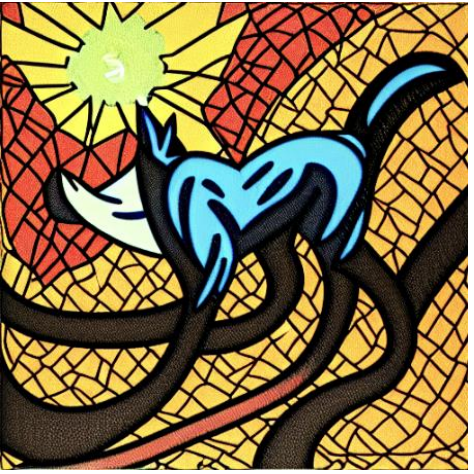


NeuralEraser

[e n s e m b l e]

Eva

Background

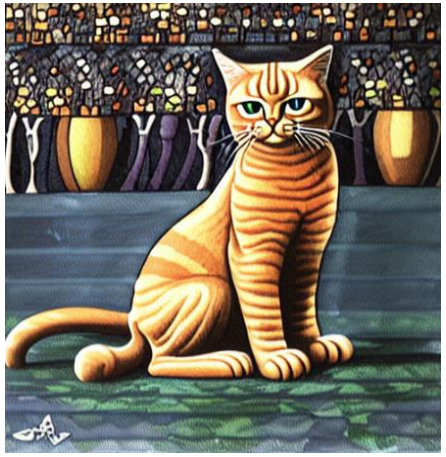
- Mitigating the spread of misinformation
- Preventing the persistence of harmful biases
- Controlling the proliferation of deepfakes

"all iPhones are set to explode in 24 hours"

Goal: find a robust & interpretable way to censor specific features in image models

"a cat in the style of an artist pencil sketch"

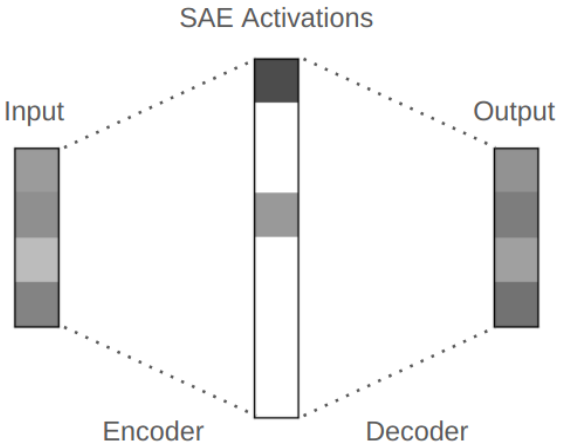
NO ARTIST PENCIL SKETCH STYLE



Previous Works

Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models

Samuel Marks* Northeastern University
Can Rager Independent
Eric J. Michaud MIT
Yonatan Belinkov Technion – IIT
David Bau Northeastern University
Aaron Mueller* Northeastern University



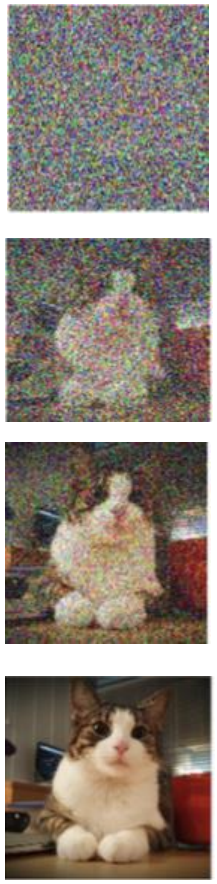
Towards Multimodal Interpretability: Learning Sparse Interpretable Features in Vision Transformers

by **hugofry** 29th Apr 2024

Unpacking SDXL Turbo: Interpreting Text-to-Image Models with Sparse Autoencoders

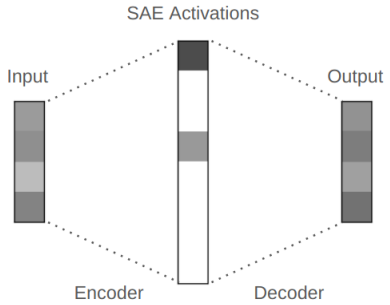
Viacheslav Surkov Chris Wendler Mikhail Terekhov Justin Deschenaux
Robert West Caglar Gulcehre
EPFL

How it works?



1 - stable diffusion model

2 - ablation



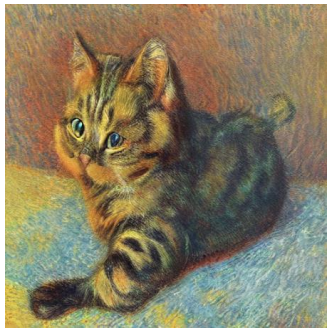
3 - SAE

Cat + Fire Style =

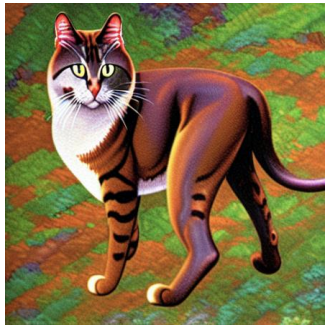
Cat + ~~Fire Style~~ =

4 - delete style

Results



"A cats image in a sponge dabbed style"



"A cats image ~~in a sponge dabbed style~~"

"A cats image in a Van Gogh style"

"A cats image ~~in a Van Gogh~~"

"A cats image in a superstring style"

"A cats image ~~in a superstring style~~"

"A cats image in a Fire style"

"A cats image ~~in a Fire style~~"