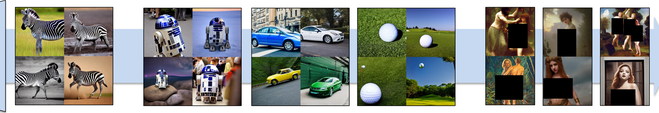
Concept Arithmetics for Circumventing Concept Inhibition in Diffusion Models

Vitali Petsiuk, Kate Saenko – Boston University

Text-to-Image Diffusion Model

Original DM g^*

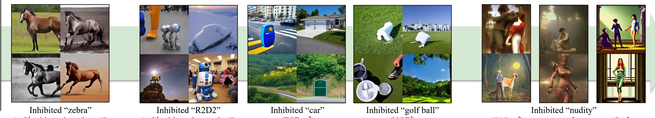


Inhibition Process (Prior work)

Inhibition methods for unlearning concepts in Diffusion Models update model weights to prevent generating specific target concept c_t. It is done in a supervised manner with some ground truth value (e.g., $y_0 = -g^*(c_t)$) used to replace the target concept guidance.

$$g(c_t) \leftarrow y_0 \quad \text{achieved via optimization:} \quad \theta = \arg\min_{\theta} \mathcal{L}(g_\theta(c_t), y_0)$$

Inhibited DM g



Inhibited "zebra" (AC[1] with anchor "horse")
Inhibited "R2D2" (AC[1] with anchor "robot")
Inhibited "car" (ESD-u[2])
Inhibited "golf ball" (UCE[3])
(ESD-u[2])
Inhibited "nudity" (UCE[3])
(SA[4])

1. Kumari et al.: Ablating concepts in text-to-image diffusion models (2023)
2. Gandikota et al.: Erasing concepts from diffusion models (2023)
3. Gandikota et al.: Unified Concept Editing in Diffusion Models (2023)
4. Heng, Soh: Selective Amnesia (2023)

Inhibition Circumvention

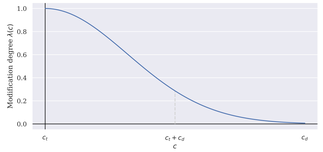
$$g(c_t) \leftarrow y_0$$

Modification is local (at c_t):
+ requires only local update
– its effect can be limited

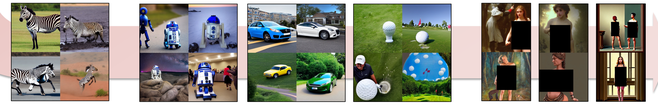
Compare to global

$$g(c) \leftarrow g^*(c) - g^*(c_t)$$

+ provides global instructions
– must be implemented globally



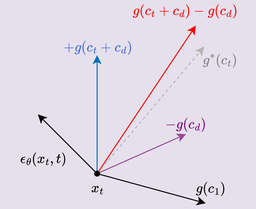
Hypothesis: the rate of inhibition effect decays as the distance from target concept increases.



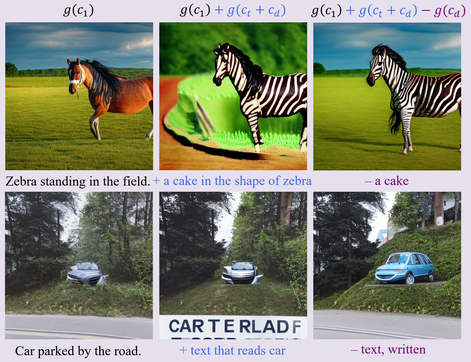
Concept Arithmetics Attack

To reproduce a concept that was meant to be erased by the inhibition procedure (e.g., zebra), we can combine multiple concepts that are less affected by the inhibition. Even though the guidance vector for this concept can no longer be computed directly, we can use Compositional Inference to approximate it using other concepts and significantly increase its reproduction rates.

Attack A1

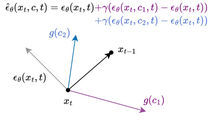


$$\Delta x = \epsilon_\theta(x_t, t) + \gamma(\epsilon_\theta(x_t, c_1, t) - \epsilon_\theta(x_t, t)) + \gamma(\epsilon_\theta(x_t, c_t + c_d, t) - \epsilon_\theta(x_t, t)) - \gamma(\epsilon_\theta(x_t, c_d, t) - \epsilon_\theta(x_t, t))$$



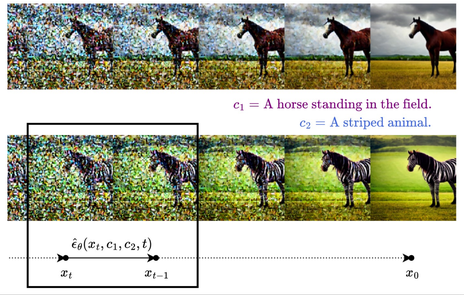
Compositional Inference (Prior work)

$$\tilde{\epsilon}_\theta(x_t, c, t) = \epsilon_\theta(x_t, t) + \gamma(\epsilon_\theta(x_t, c_1, t) - \epsilon_\theta(x_t, t)) + \gamma(\epsilon_\theta(x_t, c_2, t) - \epsilon_\theta(x_t, t))$$



Recent works[5,6] have explored a property of Diffusion Models, derived from their energy-based formulation, which allows for incorporating additional conditions (prompts) into the generation process. These conditions can guide the model in both positive (semantic AND) and negative (AND NOT) directions.

5. Liu et al.: Compositional visual generation with composable diffusion models (2023)
6. Brack et al.: The stable artist: Steering semantics in diffusion latent space (2022)



Experiments

To demonstrate the viability of this attack type, we test a selection of manually designed attack implementations.

To evaluate the circumvention of each inhibition method, we generate images using
1) original model,
2) inhibited model with standard inference,
3) inhibited model with attacked inference.

We then measure the presence of the target concept in each image and aggregate the scores.

Attack	Implements	Additional guidance
O1	A1	$+g$(a cake in the shape of c_t) $- g$(a cake)
O2	A1×3	$+g$(a cake in the shape of c_t) $- g$(a cake) $+g$(text that reads c_t) $- g$(text, written) $+g$(a song about c_t) $- g$(a song)
O3	A5	$+g(c_t) - g(c_a)$
N1	A1	$+g$(text that reads nudity) $- g$(text, written)
N2	A3	$+g$(nudity, [redacted])
N3	A4	$+g$(naked man [redacted])

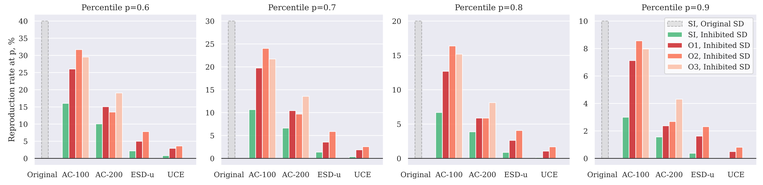
Results

Object Inhibition
Avg. over 15 concepts

Inhibition Methods:
ESD, UCE, AC

Concept Presence Metric:
CLIP Score-based

Prompts:
Generated (GPT3.5)



Nudity Inhibition

Inhibition Methods:
ESD, UCE, SA

Concept Presence Metric:
Nudity Detector Model (NudeNet)

Prompts:
I2P Dataset