Concept Unlearning in Unified-Autoregressive Models

Alexa Koeck (alexakoeck@hanyang.ac.kr) with Prof. Sungyong Baik
School of Data Science, Hanyang University

Research Motivation

Unified autoregressive (AR) models have recently created a new state of **multimodal image understanding and generation.** Text and Image generation has been the focus of unlearning approaches to **avoid copyright infringements or generating unsafe content**.

Problem

Concept erasue on LLMs has not been fully successful so far & even acceptable ones faill to account for the **dual modality** of Unified AR models. Therefore, **image-based methods can also not be applied easily** without adjustments.

This project attempts to transfer advancemts and insights from concept-erasure on text-to-image diffusion model onto unified autoregressive models.

Method

Implementing SPM (Lyu et al., 2024) on DeepSeek's open-source **Janus-Pro** model from HuggingFace.

Application Details

A **one-dimensional adapter** will be created and trained for each selected layer of the backbone Llama model implemented in Janus-Pro.

The **adapter blocks layer signals** related to the target concept as it is trained to lead the model in the **opposite gradient direction**.

The goal is to **erase the target concept without affecting other generation abilities** of the model.

Various layer combinations and loss settings were tested, as no such previous works exist.

Results

Best results come with **KL-divergence on LLM Attention Layers**. Some features remain visible.

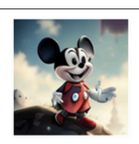
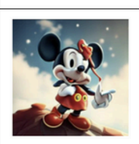
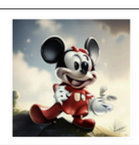
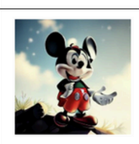
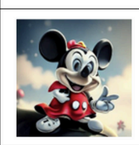
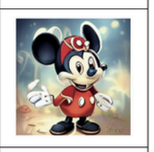
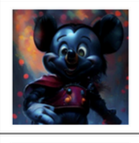
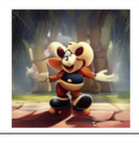
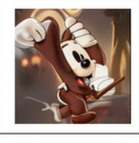
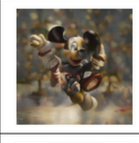
	MSE/MLP (all modules)	MSE/Attn	KL/Attn	KL/Attn+Attn	KL/Attn+MLP (all modules)	KL/Attn+MLP (7 modules)
Target Concept Epoch 0						
Target Concept Epoch 400						
Test Concept Epoch 0						
Test Concept Epoch 400						

Figure 1. Results of different loss function, modules and layer configurations for SPM on Janus-Pro. Each loss was computed on logits of each generative step of the backbone-LLM. Note: The Attn+Attn setting includes SPM membranes for attention layers in the text encoder.

Insights

Layer Selection

MLP layer filters **break overall generation abilities.** Patterns seen in Fig. 2 do not appear when using attention layers even, after further epochs.

	Target Attn+MLP	Target Attn	Test Attn+MLP	Test Attn
Output after 460 Epochs				

Figure 2. Erasing signals from MLP modules attacks overal model generation utility for both the target concept and all other general concepts.

Loss Computation

KL-divergence on logits gave the best results. This gets closer to the idea of matching logit distributions and is more suitable than MSE.

Limitations

Substitution & Anchoring

The initial **substitution mechanism** from SPM results in **no erasure effect** at all. Anchoring could not be transferred to the backbone-LLM.

	Target 40 Epochs	Target 580 Epochs	Test 40 Epochs	Test 580 Epochs
KL loss with substitute concept ("real mouse")				

Figure 3. Erasure and Generation Quality of SPM network on backbone-LLM (full attn + 7 MLP modules) with a substitute concept instead of mapping the concept to null.

Erasure-Retain Trade-off

The best erasure performance comes with text-image **alignemnt disruptions** (see Fig. 1). The test concept generation should not be affected.

Future Works

In future works, the generation quality needs to be increase to **assure effective erasure of all concept-related features.** Tests with different concept types need to be added to show generalizability, and an anchoring mechanism could help improve utility preservation. All together, this approach holds potential to provide an effective and generalizable **concept-erasure method on unified autoregressive models**.