Interpreting CLIP with Hierarchical Sparse Autoencoders

Vladimir Zaigrajew, Hubert Baniecki, Przemysław Biecek

Warsaw University of Technology

University of Warsaw

ICML International Conference On Machine Learning

Paper Code

TL;DR: Matryoshka Sparse Autoencoder is a high-performing utility tool for interpreting and controlling complex models like CLIP.

Motivation and Problem

1. **Motivation**: Understanding how complex multimodal models, like CLIP, process and represent information **is crucial** for their responsible development and deployment. **Sparse Autoencoders (SAEs)** offer a way to disentangle these complex representations into **human-interpretable features**.

2. **Problem**: Current SAEs struggle with simultaneously optimizing **both reconstruction quality and sparsity**. They often rely on activation suppression (ReLU) leading to **activation shrinkage** or **rigid sparsity constraints** (TopK). The question is, **can we train SAE withouth these limitations?**

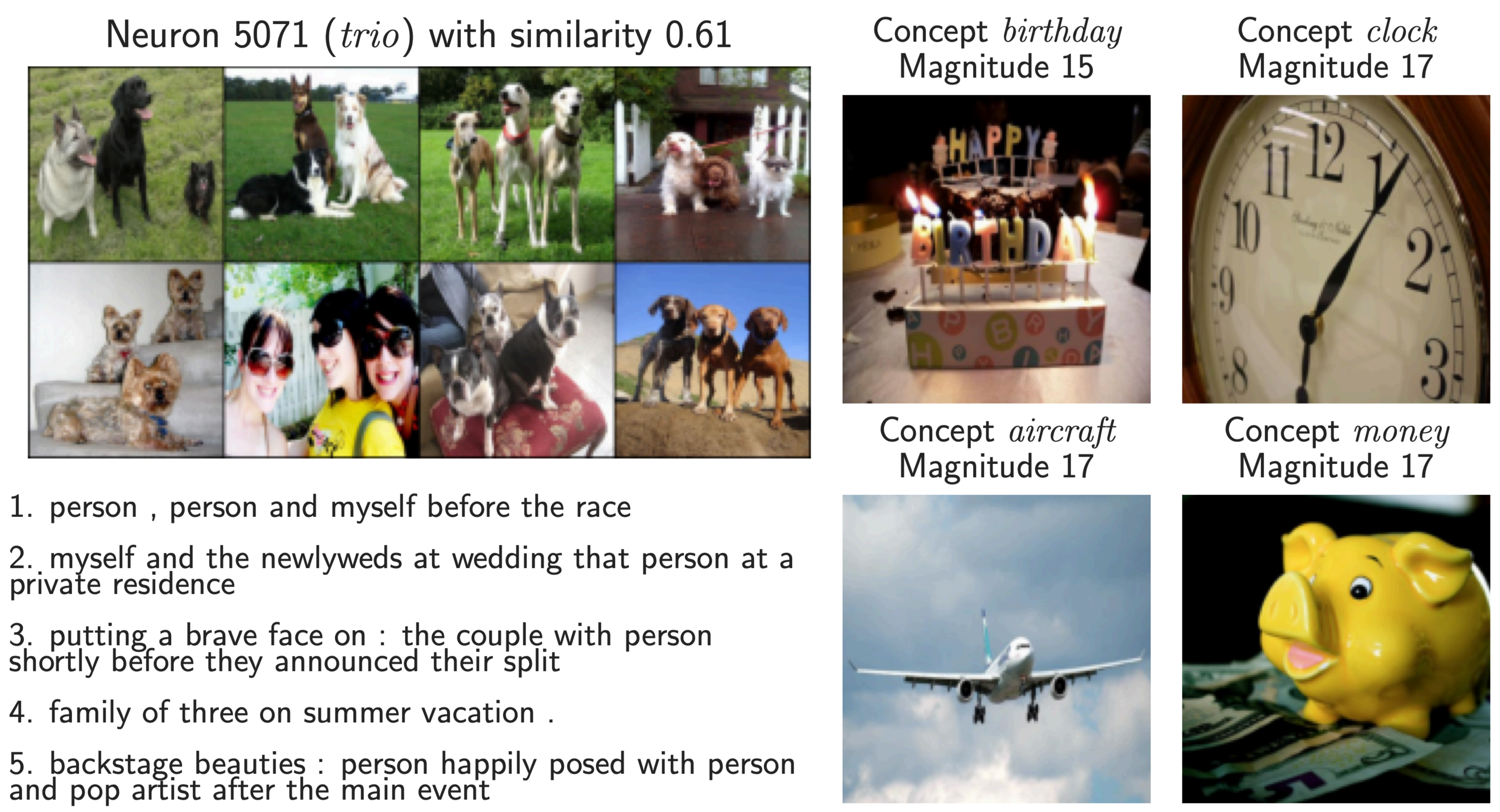
3. **Solution: Train SAE using Matryoshka representation learning!** (Kusupati et al., NeurIPS 2022)

Matryoshka Sparse Autoencoder (MSAE)

For a given input **x**, MSAE computes *h latent representations* $\mathbf{z}_i$ **during training** using a sequence of **increasing k** values $\{k_1, k_2, \ldots, k_h\}$ in *TopK function* with $|\text{TopK}_1| < |\text{TopK}_2| < \ldots < |\text{TopK}_h| \leq d$ (SAE latent size), where $\boldsymbol{\alpha}_i$ are *weighting coefficients*:

$$z_i = \text{ReLU}(\text{TopK}_i(W_{\text{enc}}(x - b_{\text{pre}}) + b_{\text{enc}})),$$
$$\hat{x}_i = W_{\text{dec}} z_i + b_{\text{pre}},$$
$$\mathcal{L}(x) := \sum_{i=1}^{h} \alpha_i \|x - \hat{x}_i\|_2^2.$$

Detected Multimodal Concepts in CLIP



Top MSAE activation samples for a given concepts

Evaluating MSAE

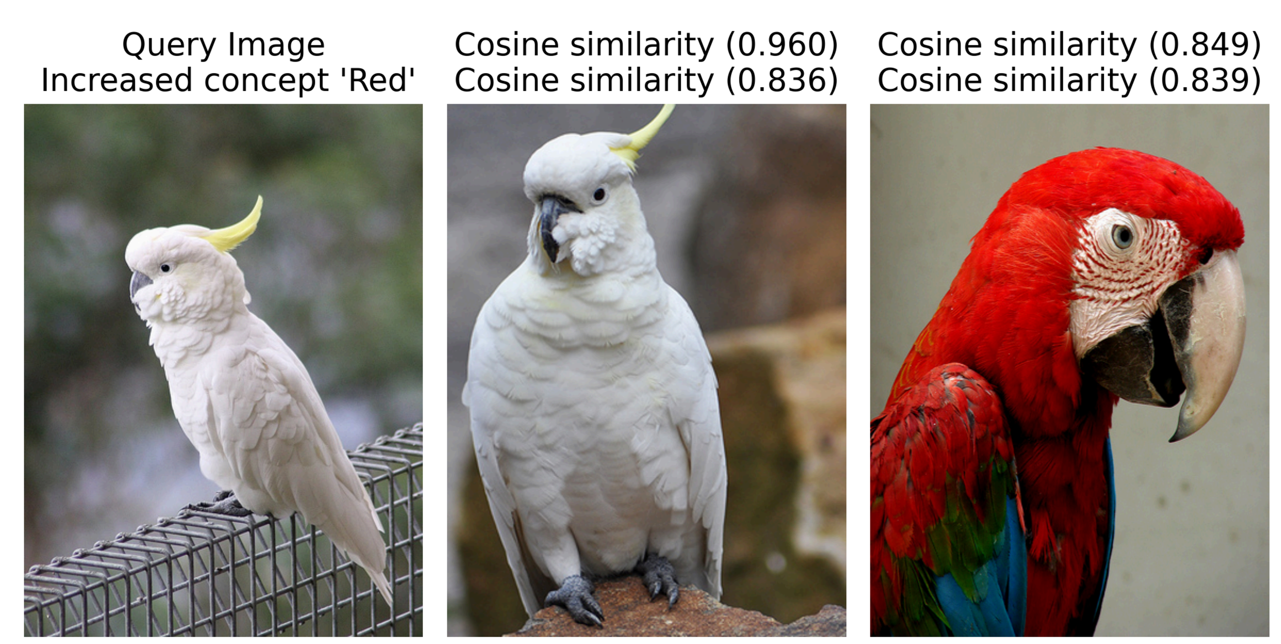
Quantitative comparison of SAE models on ImageNet-1k

Model	$L_0\uparrow$	FVU ↓	CS ↑	LP (KL) ↓	LP (Acc) ↑	CKNNA ↑	DO ↓	NDN ↓
ReLU ($\lambda = 0.03$)	$.920_{\pm.008}$	$.185_{\pm.031}$	$.928_{\pm.009}$	$50.5_{\pm77.1}$	$.977_{\pm.149}$	$.742_{\pm.005}$	.002	0(0)
ReLU ($\lambda = 0.003$)	$.649_{\pm.007}$	$.004_{\pm.000}$	$.998_{\pm.000}$	$0.66_{\pm1.03}$	$.994_{\pm.083}$	$.781_{\pm.004}$	.003	0(0)
TopK ($k = 64$)	$.950_{\pm.009}$	$.172_{\pm.026}$	$.912_{\pm.013}$	$60.1_{\pm90.8}$	$.930_{\pm.255}$	$.762_{\pm.004}$	.002	0(335)
TopK ($k = 256$)	$.900_{\pm.004}$	$.011_{\pm.003}$	$.994_{\pm.002}$	$2.71_{\pm5.40}$	$.987_{\pm.114}$	$.874_{\pm.003}$	.003	0(296)
BatchTopK ($k = 64$)	$.877_{\pm.012}$	$.162_{\pm.022}$	$.917_{\pm.011}$	$56.9_{\pm85.8}$	$.931_{\pm.253}$	$.769_{\pm.004}$	.002	0(1477)
BatchTopK ($k = 256$)	$.882_{\pm.005}$	$.010_{\pm.005}$	$.995_{\pm.002}$	$2.42_{\pm5.12}$	$.988_{\pm.108}$	$.860_{\pm.003}$	.002	3(919)
Matryoshka (RW)	$.829_{\pm.008}$	$.007_{\pm.003}$	$.997_{\pm.002}$	$3.13_{\pm7.08}$	$.987_{\pm.115}$	$.809_{\pm.002}$	.002	2(4)
Matryoshka (UW)	$.748_{\pm.006}$	$.002_{\pm.001}$	$.999_{\pm.000}$	$0.35_{\pm0.82}$	$.995_{\pm.070}$	$.848_{\pm.003}$	.002	0(22)

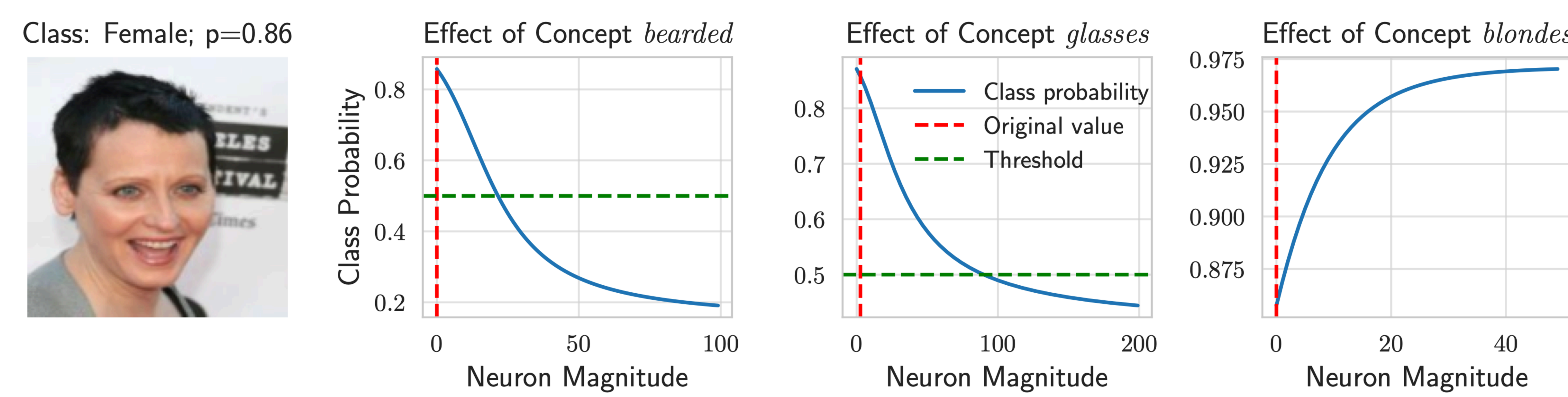
Training modality influence on MSAE performance

Matryoshka SAE variant	Language metrics on CC3M				Vision metrics on ImageNet-1k				NDN ↓
	$L_0\uparrow$	FVU ↓	CS ↑	CKNNA ↑	$L_0\uparrow$	FVU ↓	CS ↑	CKNNA ↑	
Image (RW)	$.824_{\pm.029}$	$.060_{\pm.052}$	$.971_{\pm.026}$	$.775_{\pm.001}$	$.829_{\pm.008}$	$.007_{\pm.003}$	$.997_{\pm.002}$	$.809_{\pm.002}$	4
Image (UW)	$.755_{\pm.024}$	$.026_{\pm.027}$	$.988_{\pm.012}$	$\mathbf{.790}_{\pm.002}$	$.748_{\pm.006}$	$\mathbf{.002}_{\pm.001}$	$\mathbf{.999}_{\pm.000}$	$.848_{\pm.003}$	22
Text (RW)	$\mathbf{.841}_{\pm.014}$	$.008_{\pm.003}$	$.996_{\pm.002}$	$.782_{\pm.008}$	$\mathbf{.841}_{\pm.014}$	$.008_{\pm.003}$	$.996_{\pm.002}$	$.782_{\pm.008}$	0
Text (UW)	$.791_{\pm.010}$	$\mathbf{.001}_{\pm.001}$	$\mathbf{.999}_{\pm.000}$	$.784_{\pm.007}$	$.799_{\pm.012}$	$.015_{\pm.013}$	$.993_{\pm.006}$	$\mathbf{.877}_{\pm.003}$	0

Application of MSAE to Similarity Search and Bias Validation on Dowstream Tasks



Nearest Neighbor Search with concept manipulation



Spurious correlation evaluation for gender classifcation

Contact: vladimir.zaigrajew.dokt@pw.edu.pl