Koopman Autoencoders Learn Neural Representation Dynamics

Nishant Aswani, Saif Eddin Jabari
Division of Engineering, NYU Abu Dhabi

Problem Statement and Approach

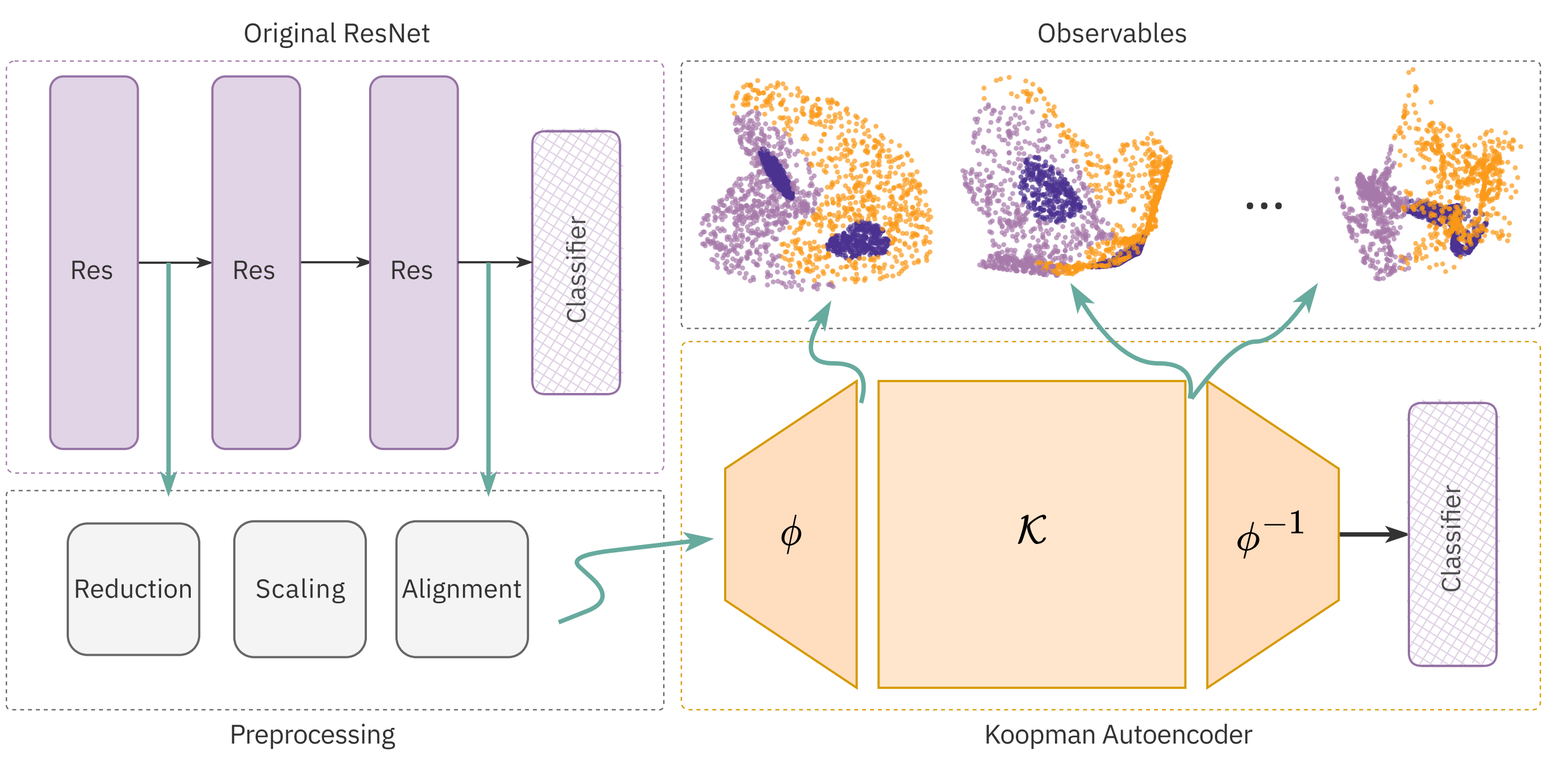
The intermediate representations of a neural network can be conceived as forming a path through high-dimensional space.

- *Can we **discover dynamics** that generate this path?*
- *Can we **edit these dynamics** to produce a different output?*

Our work relies on literature in topology, dynamics, Koopman-based approaches, and representation similarity.

To answer these questions:

- We introduce **Koopman autoencoders (KAE)** to interpolate and edit neural representations of ANNs.
- We develop an **encoder isometry** objective, which aids in preserving the original topology of neural representations in observable space.
- We demonstrate how our KAEs can be used to **edit representations** leading to fast, targeted class unlearning.



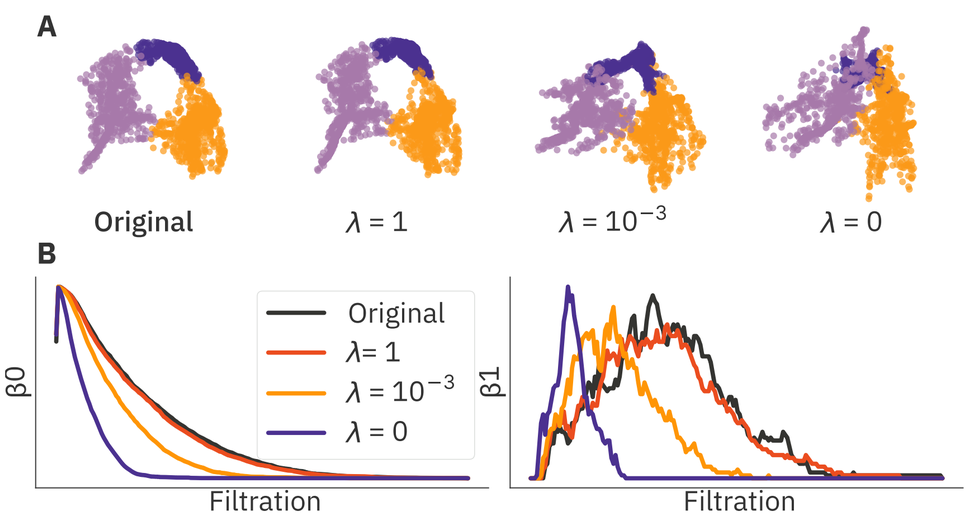
Our KAE consists of an encoder, Koopman operator, and a decoder. Trained to gradually transform an input representation to an output representation, the KAE can interpolate between them to generate a neural representation path. The model is represented as:

$$\mathbf{x}_j = \phi^{-1} \circ \mathcal{K} \circ \phi(\mathbf{x}_i),\ \forall\, i, j \in \{1, 2, \ldots, L\} : i < j$$

and trained with a weighted combination of:

$$\mathcal{L}_{\text{recon}} = \left\| \mathbf{x}_{\{i,j\}} - \phi^{-1} \circ \phi(\mathbf{x}_{\{i,j\}}) \right\|^2,$$
$$\mathcal{L}_{\text{linear}} = \left\| \phi(\mathbf{x}_j) - \mathcal{K} \circ \phi(\mathbf{x}_i) \right\|^2,$$
$$\mathcal{L}_{\text{state}} = \left\| \mathbf{x}_j - \phi^{-1} \circ \mathcal{K} \circ \phi(\mathbf{x}_i) \right\|^2,$$
$$\mathcal{L}_{\text{dist}} = \left\| \left\| \mathbf{x}_{\{i,j\}} \right\|^2 - \left\| \phi(\mathbf{x}_{\{i,j\}}) \right\|^2 \right\|^2,$$

Encoder Isometry



We demonstrate that the most strongly penalized encoder (red) exhibits the closest topological similarity to the original model (black). These results indicate that increasing the encoder isometry penalty leads to more topologically faithful representations in observable space. As a result, we expect that topological edits in the observable space will also be reflected in the state space.

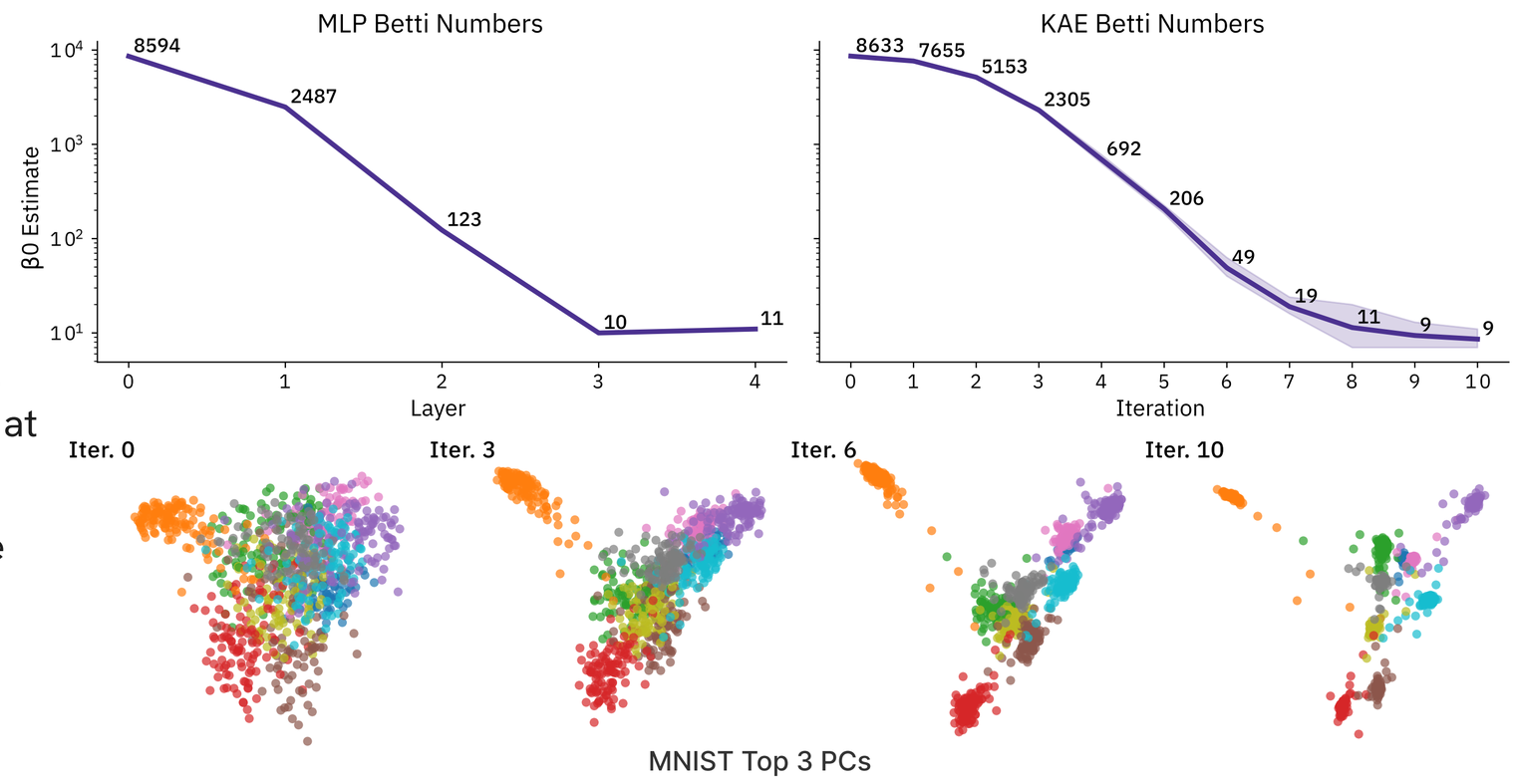
See our KAEs in action!

Matching Accuracy and Simplifying Topology

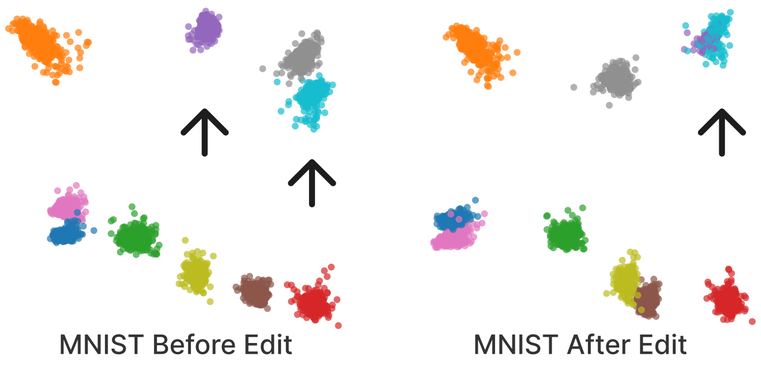
Dataset	MLP % Acc.	KAE % Acc.
Yin-Yang	99.31	98.75 (0.15)
MNIST	99.03	98.53 (0.04)

Class accuracy provides a measure of surrogate quality. Our KAEs are able to faithfully produce the penultimate layer representations for both datasets. Additionally, our KAEs naturally simplify in topology at every step.

We hypothesize that the KAE dynamics can be made more faithful to the original residual network by regularizing the KAE's intermediate representations.



Model Editing



Dataset	Target Class	Edited Acc. (StDev.)
Yin-Yang	Class 0 (Yin)	98.78 (1.18) → 85.01 (1.90)
	Class 1 (Yang)	98.27 (0.21) → 78.88 (8.53)
	Class 2 (Dots)	99.97 (0.05) → 62.52 (1.35)
MNIST	Class 1	99.23 (0.04) → 0.0 (0.0)
	Class 4	98.29 (0.08) → 0.0 (0.0)
	Class 7	98.01 (0.18) → 0.0 (0.0)

In observable space, assuming simple topology, we can identify unwanted outputs and their corresponding inputs. With a model editing algorithm, we can learn an edited linear operator which generates an updated representation—sans the unwanted outputs.