Weight Decay Shapes Representation Geometry: Towards a More Nuanced Understanding of Sparse Autoencoders in Vision Transformers

Nina Burdorf, Christian M. Adriano
Hasso Plattner Institute, Potsdam, Germany

HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

ICML International Conference On Machine Learning

PAPER

1 TL;DR

MAIN TAKEAWAYS

- The way a ViT is trained influences the representations that SAEs learn
- Weight decay is the strongest training factor we observed affecting SAE behavior
- Below a small weight-decay threshold, SAE feature usage collapses into a few repeatedly reused features
- Our findings suggest that both the underlying ViT representations and task complexity shape SAE behavior

2 Background and Motivation

- Mechanistic interpretability often studies already-trained models
- However, previous work suggests that training choices influence representation geometry, raising an important question: Do training decisions affect how well Sparse Autoencoders can recover interpretable features?
- To investigate this, we train Vision Transformers (ViTs) across different optimization settings and analyzed the SAEs trained on their internal representations

3 Experimental Design

- 64-model sweep across six ViT hyperparameters
- Matched weight-decay sweep to isolate the effect of regularization
- SAEs trained on multiple ViT residual layers
- Evaluation: CE recovery, monosemanticity, dead features, representation geometry, and feature usage

4 Key Findings

Weight Decay is the Most Consistently Associated Hyperparameter

- Significant **association with dead-feature counts** across all analyzed layers
- **Other hyperparameters** showed **no** or layer-specific **effects**
- Classification accuracy remained comparable across matched weight-decay models

Better SAE Quality Aligns with Better Representation Structure

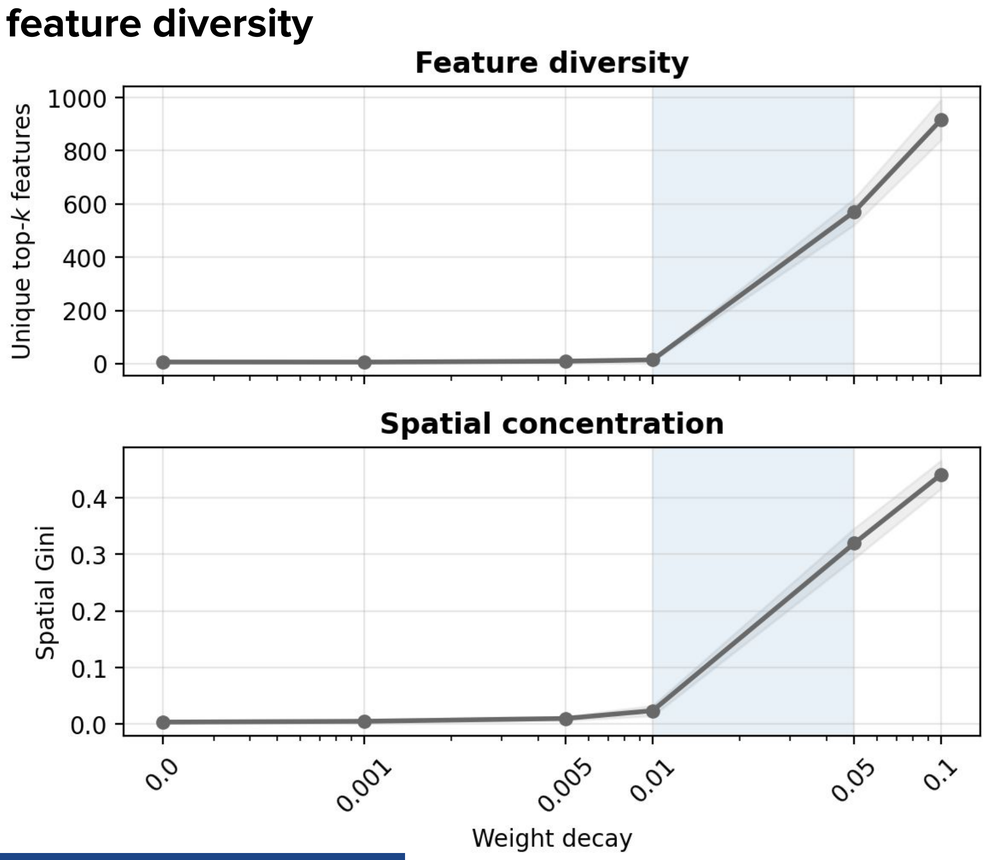
- Higher **monosemanticity → improved CE recovery**
- **Fewer dead features → improved reconstruction**
- **Strongest** relationships appear in **deeper ViT layers**

	Monosemanticity vs. CE Recovered		Dead Features vs. CE Recovered	
Layer	τ	p	τ	p
3	-0.1538	0.0725	0.0154	0.8575
5	-0.1468	0.0864	0.0095	0.9122
7	0.0595	0.4869	-0.1778	0.0460
8	0.1518	0.0763	-0.2866	0.0015
9	**0.1944**	**0.0231**	**-0.3009**	**0.0009**
10	**0.2877**	**0.0008**	**-0.3964**	$\mathbf{1.0 \times 10^{-5}}$
11	**0.3333**	**0.0001**	**-0.7445**	$\mathbf{6.3 \times 10^{-17}}$

Mann–Whitney U tests for dead-neuron counts across weight decay settings.

Low Weight Decay Leads to Feature Collapse

- **Below approximately wd ≈ 0.01 the same small feature set repeatedly activates**
- Higher **weight decay** produces substantially greater **feature diversity**



Task Complexity Matters

- **More classes reduce SAE reconstruction quality**
- Indicates that training configuration is only one factor affecting sparse decomposition

Samples per class	Kendall's τ	p
200	**0.810**	**0.0107**
400	**0.905**	**0.0028**
600	0.619	0.0690
800	**0.810**	**0.0107**
1000	**0.810**	**0.0107**
1200	**0.905**	**0.0028**

Kendall's τ correlations between number of classes and average reconstruction CE across different sample-size settings.

5 Conclusion

Discussion

- Training choices of the underlying model may influence the representations that SAEs learn
- Small changes in regularization can move models between feature-usage regimes
- Training may therefore be considered when interpreting SAE results, not only the SAE architecture itself

Limitations

- Results are limited to ViT-Tiny models
- Experiments use traffic-sign datasets only
- Findings are correlational and do not establish causality
- Monosemanticity depends on a CLIP-based evaluation metric

Future Work

- Evaluate whether the observed effects extend beyond SAEs to other mechanistic interpretability methods
- Develop training objectives that jointly optimize model performance and interpretability

nina.burdorf@student.hpi.uni-potsdam.de