Supplement for: Probing the Representational Power of Sparse Autoencoders in Vision Models

Matthew L. Olson[1], Musashi Hinck[1], Neale Ratzlaff[1], Changbai Li[2],
Phillip Howard[1], Vasudev Lal[1], Shao-Yen Tseng[1]
[1]Intel Labs, Santa Clara, CA, USA [2]Oregon State University, Corvallis, OR, USA
[1]{matthew.lyle.olson, musashi.hinck, neale.ratzlaff,phillip.r.howard,
vasudev.lal, shao-yen.tseng}@intel.com, [2]lc@oregonstate.edu

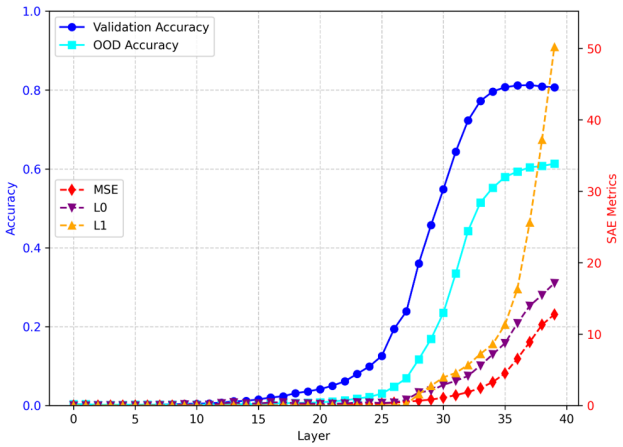

Figure 1. Results of training a ReLU SAE on every layer of DINOv2.

1. SAEs Layer-by-layer

In figure 1 we find early layers of DINOv2's CLS representation contain no information. This suggests SAEs can be used to identify meaning in a given model's token representations simply by measuring the unsupervised SAE metrics.

2. Steering Stable Diffusion3.5

In figure 2, we show examples of steering on StableDiffusion3.5 [3]. We train an SAE on the full text output of the three text encoder models, treating each positional representation as independent. We found some highly activate text embeddings, then used those to steer the model at test time. While these activations are consistent, many of the learned SAE activations were not semantically meaningful. The exact prompts for the starting images are:

1. *close up shot of Supermans cape flapping in the wind, glowing neon, in the style of Yoji Shinkawa, wide shot, dark and gritty Superman film, visual striking colors, neon demon vibes, epic, Superman is standing on a tall statue, his red cape is flapping in the night sky, shot on Afga Vista 400, natural lighting*
2. *A fit person wearing a suit and tie sitting at the desk. Head of a Koala taken with a Canon in hyperrealistic 4k on complete Black background.*
3. *back to school cute black girl cartoon sticker*

3. Full SAE results.

In figure 3 we show detailed results for the full hyperparameter sweep for different SAEs, Vision Models, and Expansion sizes. In figure 4 we show a detailed analysis of just ResNet [2], finding extreme consistency across models and expansion sizes. In figure 5 we conduct an experiment where we fit skip the activation function for the SAE when fitting a linear layer for classification. By ignoring the activation function, SAEs trained with stricter L_1 penalties are able to achive high accuracy.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[3] Stability AI. Introducing stable diffusion 3.5, 2024. Updated October 29, 2024. 1

Figure 2. **Steering StableDiffusion 3.5**. We steer the generation of three starting images using three learned SAE features. While SD3.5 is challenging to steer, using an SAE to steer is technically possible.

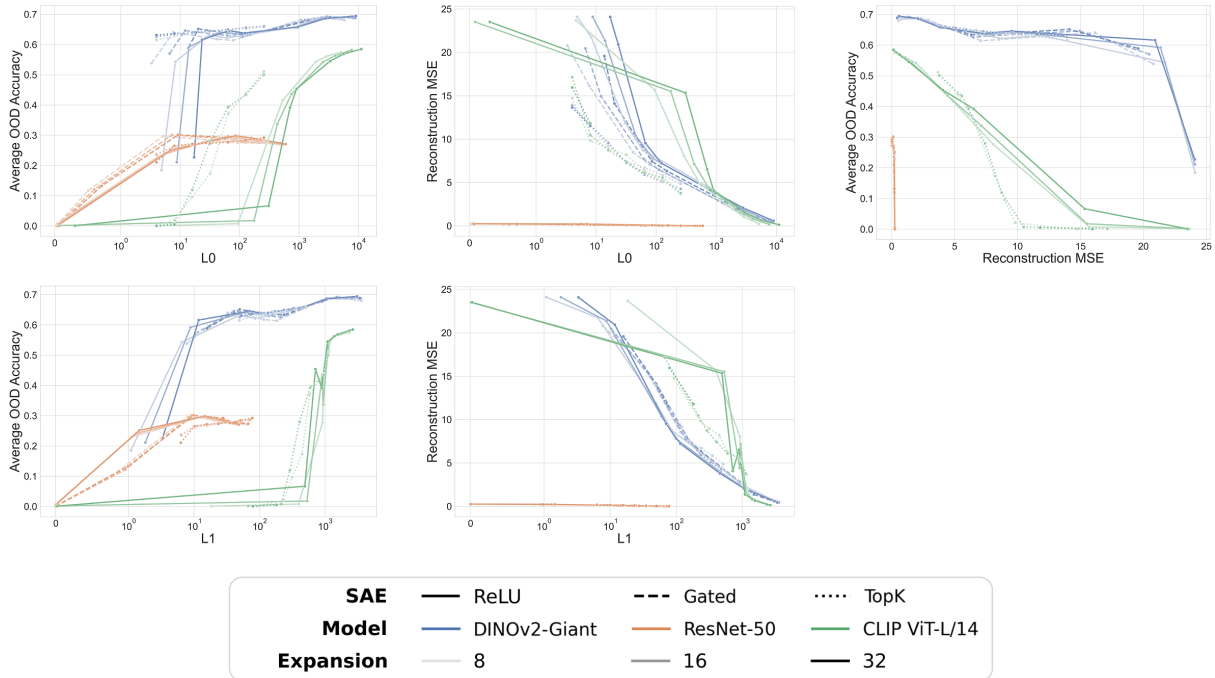

Figure 3. An expanded version of Table 1. We show OOD accuracy and Reconstruction error versus sparsity metrics.

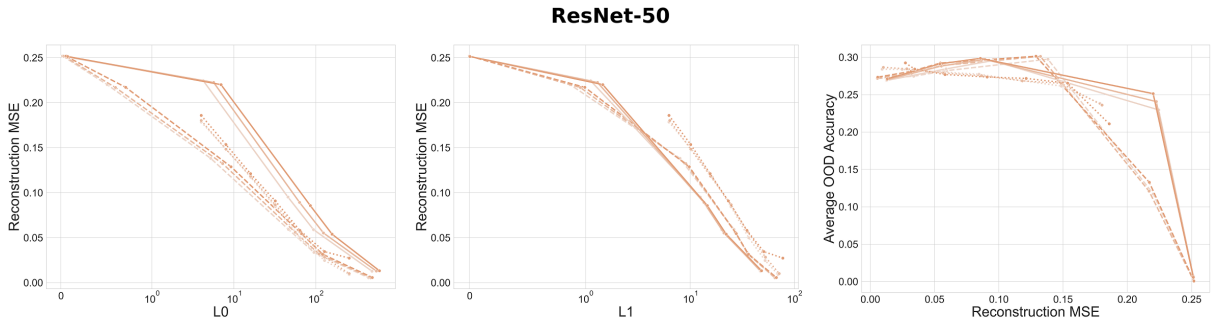

Figure 4. A detailed analysis of SAEs trained on ResNet. Across different hyper-parameters and SAE types, we find very consistent trends.

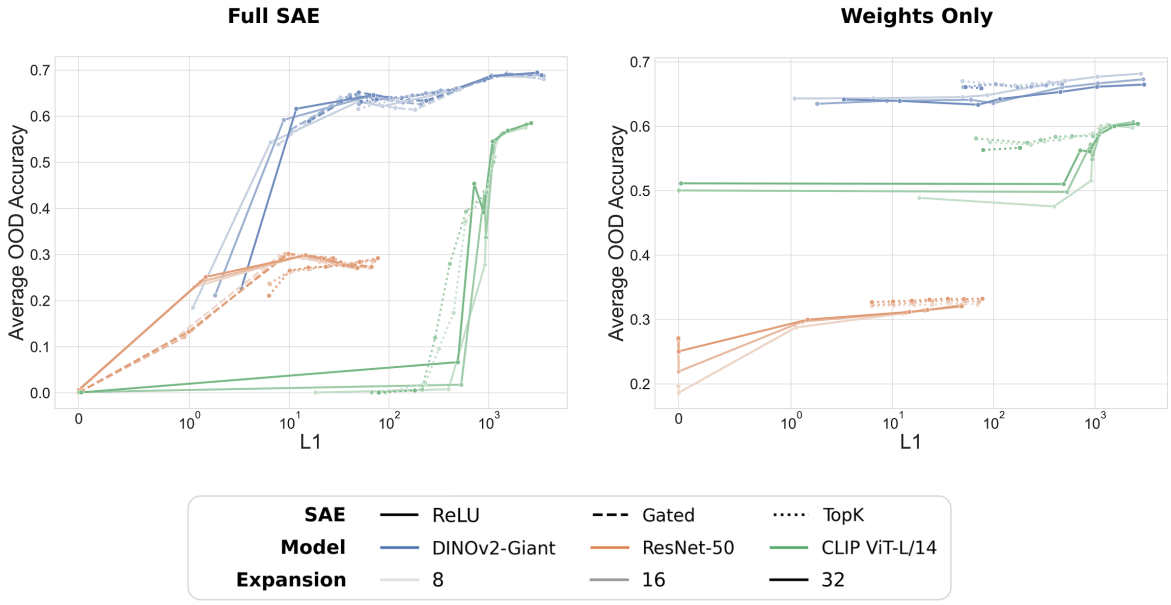

Figure 5. An experiment to measure how ignoring the SAE activations effect downstream performance on fitting a linear layer to classify ImageNet [1]. We find ignoring the activations improves performance for models with original high sparsity.