Mammo-SAE: Interpreting Breast Cancer Concept Learning with Sparse Autoencoders

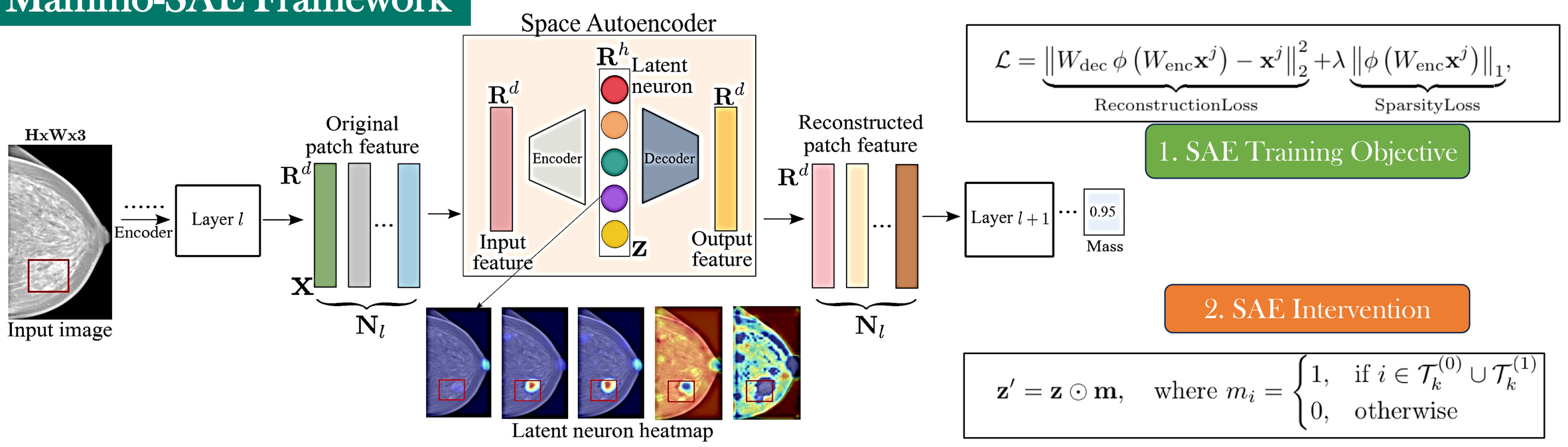
Author: Krishna Kanth Nakka

Contributions

TLDR: We train a Sparse Autoencoder on vision features from the foundation models to identify interpretable neurons responsible for breast-concept predictions, uncovering model behaviour and improving explainability.

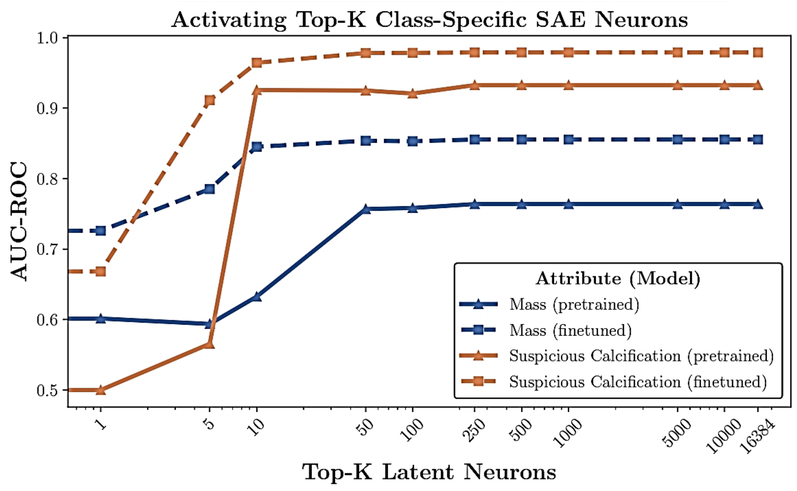
- We introduce **Mammo-SAE**, trained on the Mammo-CLIP breast foundation model, and show that the neurons in the SAE latent space are human-interpretable.
- We intervene on Mammo-SAE latent neurons to understand model behavior in downstream predictions.
- We observe that the top-activated latent SAE neurons align with true regions in the fine-tuned models, thereby uncovering the reasons behind performance gains.

Mammo-SAE Framework

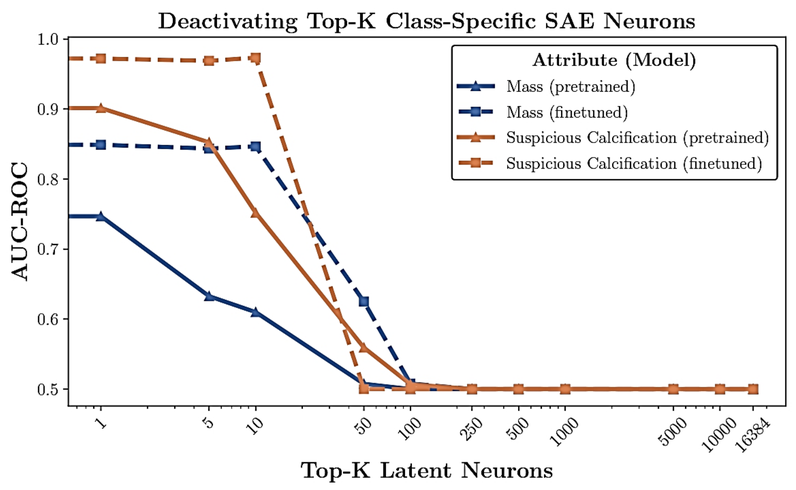


Intervention on SAE Neurons

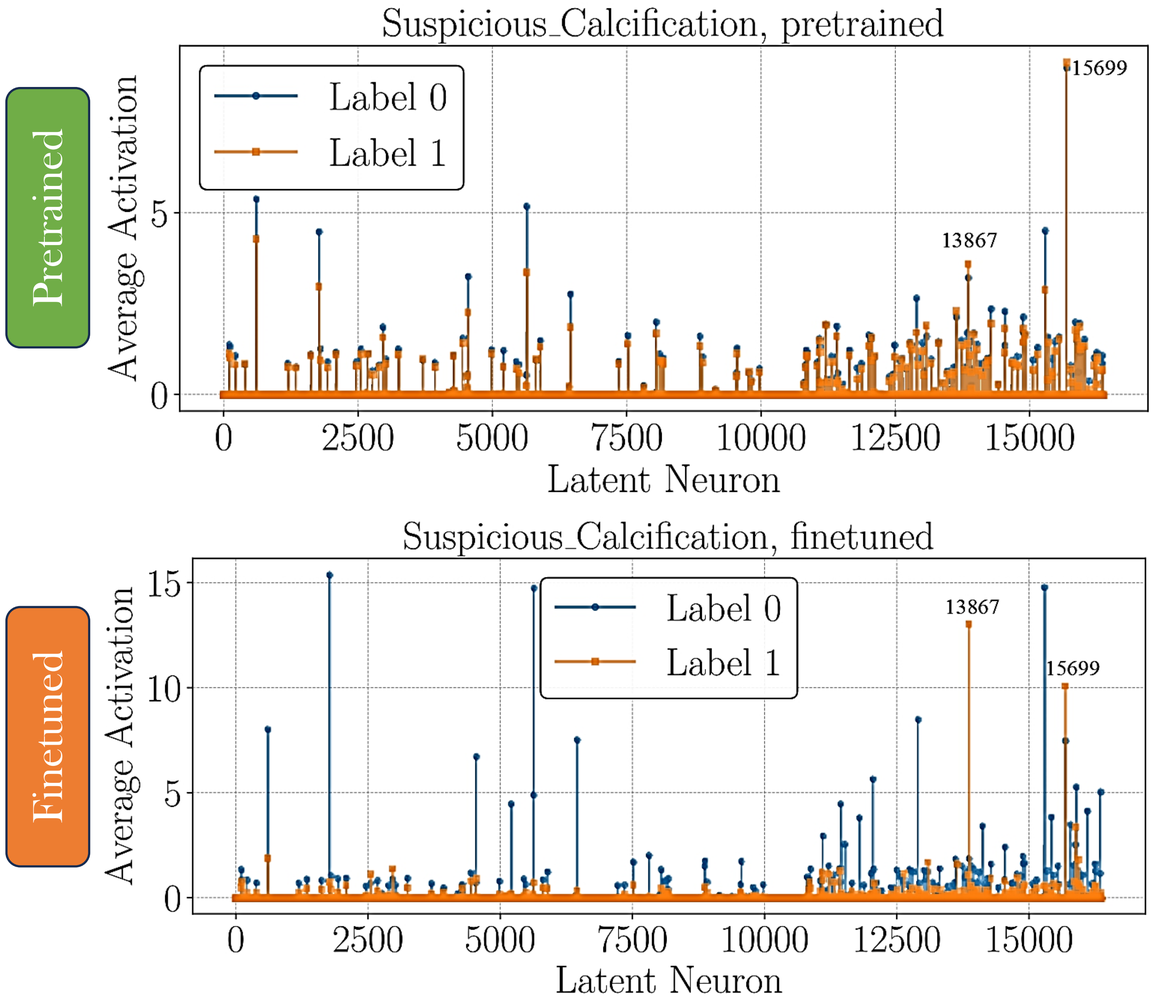
Top-k SAE Neuron Activation



Top-k SAE Neuron Deactivation



Pretrained vs Finetuned: SAE Activations



SAE Neuron visualization