
Experimenting with Universal Sparse Autoencoders to Analyze Multiple Models Internal Representation Alignment

Mohammad Salman Ahmed

1. Introduction to Mechanistic Interpretability

Mechanistic interpretability is a field that challenges the "black-box" theory surrounding the functionality of neural networks. It seeks to decode how these networks really work and provide human-understandable reasoning for their decisions. Instead of treating neural networks as opaque systems, mechanistic interpretability looks inside the models to identify specific subgraphs or circuits that implement distinct behaviors.

One fundamental barrier to interpretability is polysemasity—a phenomenon where a model with fewer dimensions is forced to learn multiple features. As a result, each neuron may encode information for multiple features at once, even when those features appear to be irrelevant to each other. This phenomenon complicates our ability to directly map activations to specific features.

A potential resolution to polysemasity is the concept of "directions in activation space." The key insight here is that features are not represented by individual neurons but by directions in the activation space. An activation vector is an n -dimensional vector, where n is the number of neurons in the layer. Each activation is a linear combination of feature directions that lie within the same space but point in different directions.

In this experiment, we aim to discover these feature directions from the activations of two pre-trained models, namely ResNet and Vision Transformer (ViT), both trained on ImageNet. Our goal is not only to discover the features these models learn but also to evaluate the Platonic Representation Hypothesis—the idea that different models converge to the same fundamental features, regardless of architecture.

Implementation for this experiment can be found on my github with the given link: https://github.com/SalmanDeAnalyst/mechanistic_interpretability_using_usae.git

1.1. Methodology

For this experiment, we selected two models: ResNet and ViT, both pretrained on ImageNet. The dataset used for evaluation is CIFAR-10. We begin by collecting the activa-

tions from both models using a set of 10,000 images and then form activation pairs from the two models. We then train a Universal Sparse Autoencoder (USAE) with shared latent variables (denoted as z) of dimensions 8190 and 2042. The model is trained for 100 epochs, during which we randomly select the source model (ResNet or ViT) and attempt to reconstruct the activations of both models from the shared z .

To introduce sparsity into the model, we use the TopK algorithm with $k = 64$, ensuring that only the most relevant activations are considered. After training, we compute a squared confusion matrix to analyze how well the model can cross-reconstruct the activations from the learned shared z .

Next, we evaluate the feature extraction (FE) performance to determine if the extracted features are universal (common to both models) or abstract (specific to each model's internal representations). Once we identify the universal features, we attempt to decode whether these features represent similar concepts across both models or whether they differ, using coordinated activation maximization.

Finally, we assess whether individual Sparse Autoencoders (SAEs) perform better at feature learning than the Universal SAEs.

1.2. Analyzing USAE Training with Shared Z

In our USAE experiments, we found that cross-model reconstruction quality consistently improved when we used a smaller shared code dimension, higher sparsity, and more training epochs. Initially, we trained the model with a large shared space (z -dim = 8168) and Top-K = 32. This configuration produced weak cross-reconstruction scores (ResNet \rightarrow ViT 0.32, ViT \rightarrow ResNet 0.36), indicating limited universality.

After reducing the shared dimension to 2042 and increasing sparsity to $K = 64$, cross-reconstruction improved significantly. As seen in Figure 1, the off-diagonal R^2 scores increased (ResNet \rightarrow ViT = 0.360, ViT \rightarrow ResNet = 0.417), showing stronger universal decoding. Model-specific reconstruction also improved (0.729 and 0.720 respectively).

This behavior is expected because our backbone models have relatively small activation sizes (512 for ResNet, 768 for ViT). Expanding these to an excessively large shared space (8168) introduces redundancy, making it harder for the encoders and decoders to learn meaningful universal features. Additionally, using a very strict sparsity level ($K = 32$) further constrained the representation capacity.

Given these observations, training for more epochs (e.g., 200+) would likely yield even better universality and stronger off-diagonal R^2 values.

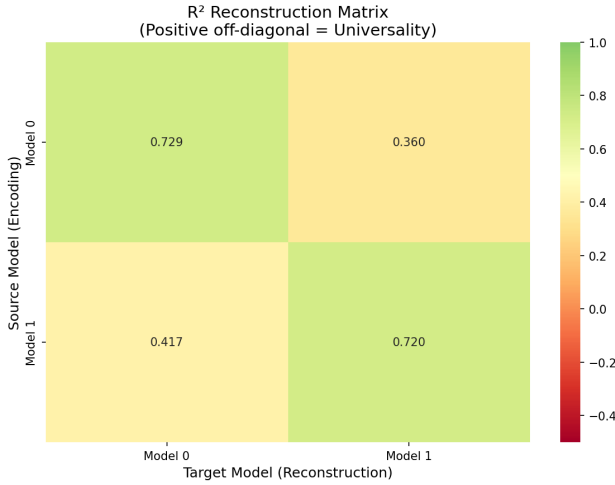


Figure 1. Reconstruction Confusion Matrix

2. Quantifying Universality

The firing entropy plot shows a strongly bimodal pattern. Almost 11 active features cluster near F E 1.0 as shown in figure 2, meaning they fire almost equally for both models and therefore represent universal shared structure learned by the USAE. A smaller set of features appears near F E 0, indicating model-specific artifacts that only activate for one model’s representations. This separation is expected: the sparse shared space discovers a core set of features that generalize across architectures, while still preserving a few idiosyncratic signals tied to each model’s internal processing.

Overall, the presence of a dominant high-entropy mode is a positive sign our dictionary contains genuinely universal features rather than being dominated by single-model noise.

2.1. Visualizing Consensus using Coordinated Activation Maximization (CAM)

The CAM results for Feature 273, as shown in figure 3, show that both models converge toward high-frequency,

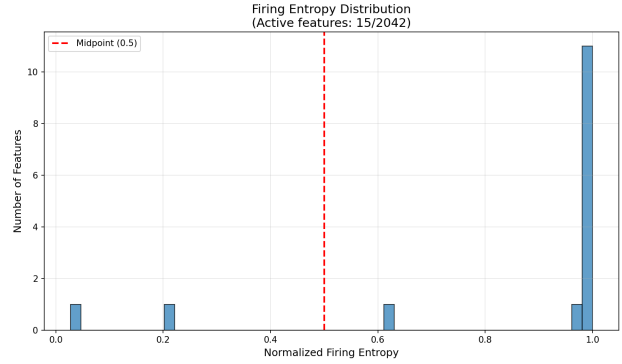


Figure 2. Firing Entropy metric for K features

texture-like patterns, consistent with its Firing Entropy = 1.0 (fully universal). Even though the exact pixel structures differ—ResNet produces slightly more structured, edge-like textures while ViT produces a denser, more isotropic pattern—both clearly optimize toward a shared abstract concept: a highly textured, multi-orientation stimulus that strongly activates this universal feature.

The optimization curves reinforce this: both models steadily increase activation for the same feature, with ViT achieving a higher final activation but following a similar trajectory. This indicates that both architectures agree on the underlying representation even if their visualization manifolds express it differently.

Overall, Feature 273 is a strong example of universality with mild representational divergence: the concept is shared, but each model realizes it in its own architectural style.

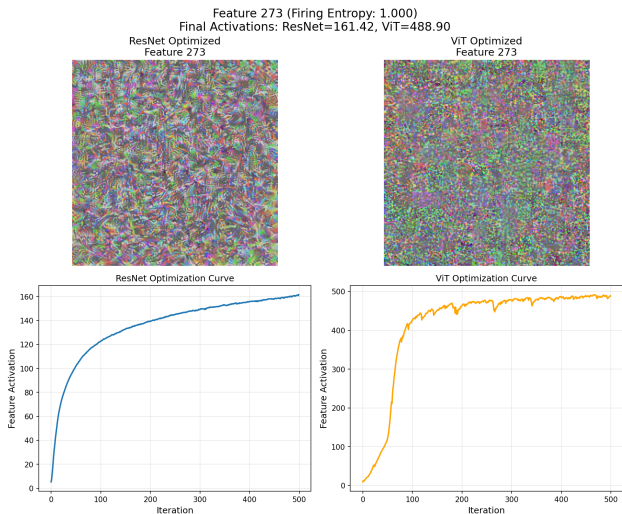


Figure 3. CAM results for Feature 273

2.2. USAE vs Individual SAE

The Alignment Tax Analysis directly addresses whether forcing features to be shared degrades their quality by comparing the Self-Reconstruction R^2 Score between Independent Sparse Autoencoders (SAEs) and the Universal Sparse Autoencoder (USAE). A drop in the R^2 score signifies an Alignment Tax—a loss in the ability to reconstruct the original hidden state. Our experiment revealed a strong, model-dependent trade-off. The ResNet paid a significant tax, with its R^2 score dropping from 0.818 (Independent SAE) to 0.743 (USAE), representing a $\sim 9\%$ loss in fidelity. This high tax suggests the ResNet’s internal features struggled to compromise, potentially making the interpretability gain difficult to justify due to the substantial sacrifice in representational quality.

In stark contrast, the ViT model incurred only a minimal Alignment Tax, with its R^2 score dropping negligibly from 0.754 (Independent SAE) to 0.750 (USAE). This indicates that the ViT’s native feature space was either inherently better suited to the universal basis or that the shared features aligned more naturally with its patch-based representations. Therefore, for the ViT, the benefit of having a truly shared, universal set of features across architectures is easily justified, as the gain in interpretability is achieved with almost no performance degradation. The overall conclusion is that while USAE successfully creates a universal basis, the Alignment Tax is a real cost that must be evaluated case-by-case.

universality. The firing-entropy analysis revealed a clear separation between universal and model-specific features, with several features achieving near-maximal entropy and behaving consistently across models. CAM visualizations further validated this: both models optimized toward similar high-entropy concepts, albeit with architecture-specific texture differences. Overall, these results demonstrate that universal features do emerge even between models with very different inductive biases.

Possible improvements include training for more epochs, exploring smoother sparsity (e.g., soft-top-k or L1 annealing), learning adaptive Z-dimensions per feature, adding cross-model contrastive losses, and scaling to deeper layers or more diverse architectures to strengthen universality.

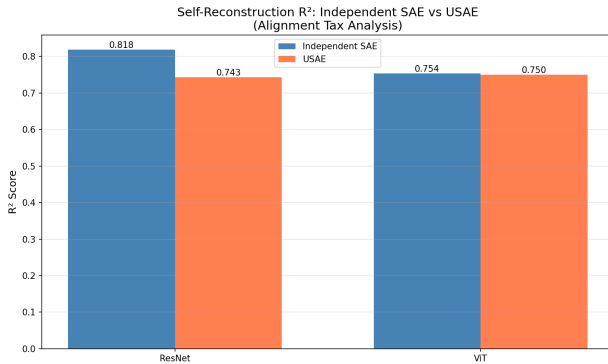


Figure 4. USAE vs Individual SAE Reconstruction score

2.3. Conclusion

Our experiments show that the Universal Sparse Autoencoder (USAE) successfully learns a shared latent space that captures cross-model structure between ResNet and ViT. Reducing the Z-dimension and relaxing sparsity (higher Top-K) significantly improved cross-reconstruction, confirming that overly large or overly sparse bottlenecks hurt