Interpreting Visual Transformers with Sparse Autoencoders

Team members: Anton Korznikov Mikhail Kuznetsov Yurii Potapov

Project supervisor: Ruslan Rakhimov

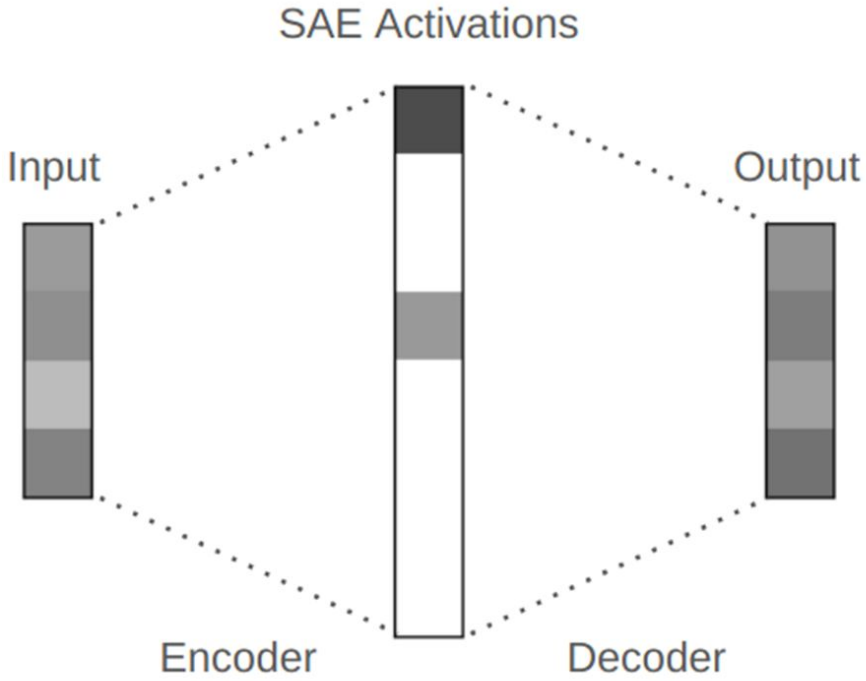
Problem statement

The main goal of our project was to explore the possibility of interpreting intermediate or output activations of visual transformers using Sparse Autoencoders, as is done in NLP research where they successfully interpret LLMs.

Objectives of the project

- Train SAE (Sparse Autoencoder) on the output activations of the Theia ViT model and interpret the resulting "features"
- Take the trained SAE and deploy it as a head on top of the Theia model. Test the SAE-headed model on CortextBench
- Investigate (including via SAE) how the DepthAnything model encodes depth predictions within its internal activations.

What are Sparse Autoencoders?



superposition hypothesis: $\mathbf{x} \approx \mathbf{x}_0 + \sum_{i=1}^{M} f_i(\mathbf{x})\mathbf{d}_i,$

embedding vector of some token on 8th layer residual stream of transformer

SAE architecture and training loss:

$$\mathbf{f}(\mathbf{x}) := \mathrm{ReLU}\left(\mathbf{W}_{\mathrm{enc}}\left(\mathbf{x} - \mathbf{b}_{\mathrm{dec}}\right) + \mathbf{b}_{\mathrm{enc}}\right)$$

$$\hat{\mathbf{x}}(\mathbf{f}) := \mathbf{W}_{\mathrm{dec}}\mathbf{f} + \mathbf{b}_{\mathrm{dec}}$$

$$\mathcal{L}(\mathbf{x}) := \|\mathbf{x} - \hat{\mathbf{x}}(\mathbf{f}(\mathbf{x}))\|_2^2 + \lambda \|\mathbf{f}(\mathbf{x})\|_1 .$$

Existing solutions and novelty of our project

Despite the surge in popularity of Sparse Autoencoders (SAEs) for interpreting the inner workings of LLMs, there are very few articles exploring the application of SAEs to visual transformers.

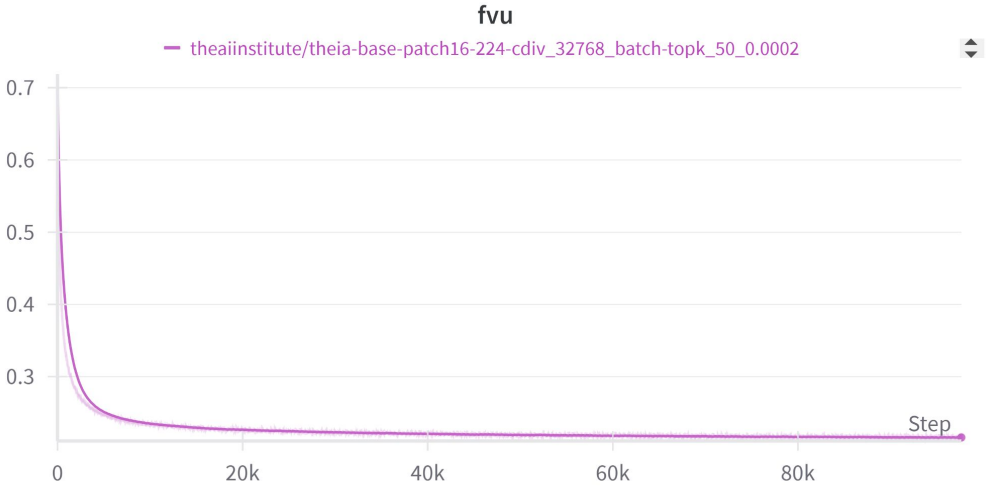
Existing studies are limited to investigating SAEs for interpreting features of models like CLIP and do not extend their research further.

The novel aspects of our project include:

- First successful training of SAEs on spatial tokens (instead of the [cls] token used in all prior work).
- Exploration of robustness of visual encoders using the Theia model and the CortexBench.
- SAE-driven investigation of how the DepthAnything model encodes depth-prediction features.

Training SAE

- We had to write the entire code for training SAEs on visual transformers from scratch.
- To avoid overloading the RAM, we had to develop a separate buffer module to store activations.
- We trained SAE on Theia-base output activations on ImageNet images.
- We used the BatchTopK architecture for the SAE and set L0 = 50. Dictionary size was 2**15 latents.
- The SAE was trained on 400 million tokens for 8 hours on a single A100.



Main metric of success training was Fraction of Variance Unexplained:

$$\mathrm{FVU} = \frac{\mathrm{MSE}(f)}{\mathrm{var}[Y]}$$

SAE latent № 5776. Top 5 activations.

SAE latent № 4714. Top 4 activations.

SAE latent № 9073. Top 4 activations.

SAE latent № 12714. Top 5 activations.

SAE latent № 1117. Top 6 activations.