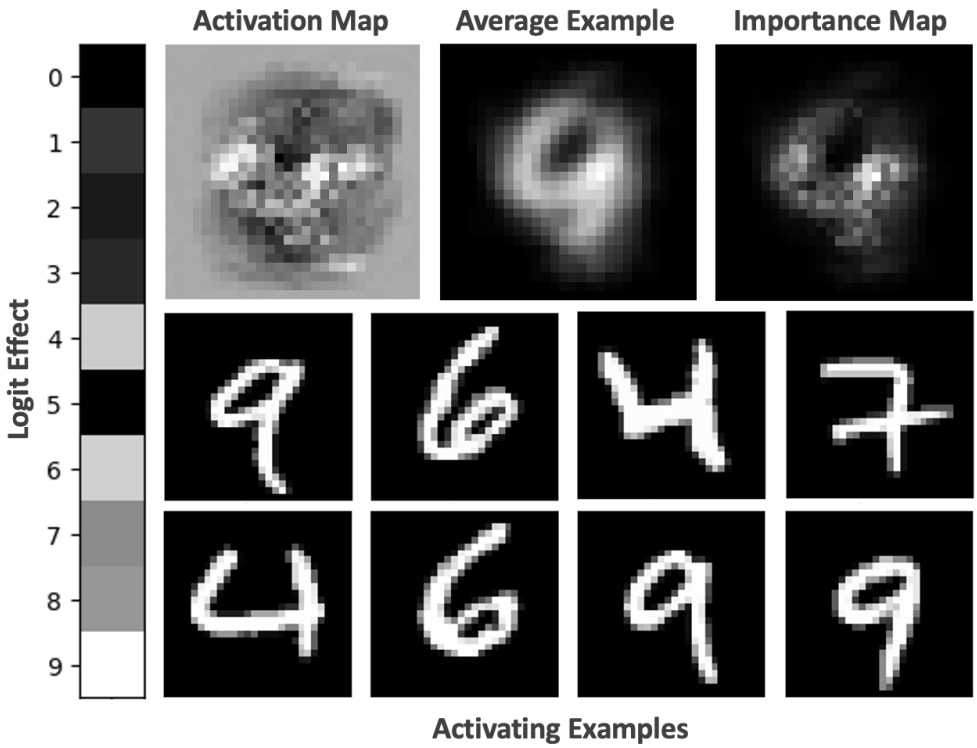

Figure 5. Visualisations for an SAE neuron. The activation map shows the maximum magnitude of neuron activation for each pixel in the input, and the importance map is the average dataset example scaled by the activation map.

ples and clusters of examples. By creating feature clusters to separate a neuron's dataset examples into their distinct behaviours we make it much easier to interpret the neuron, and also reduce the risk of running into the interpretability illusion by making it feasible to collect and summarise a wide variety of examples from across the activation spectrum. As dataset examples are an input into some automated interpretability techniques, applying feature clustering first could improve the results of these tools as well.

However, we note that our work doesn't consider features that may only emerge when considering several neurons together, which is a significant limitation. Future work could investigate using neuron embeddings and feature clusters in circuit analysis, or even look to extend the representation to multiple neurons. For example, it could be easier to understand how neurons co-activate to compensate for superposition by measuring which sub-neuron features activate together, rather than analysing neuron activation correlations directly as polysemanticity makes this very challenging. Additionally, in language models we only use the pre-MLP embedding of the token with the highest neuron activation. Combining the embeddings of multiple tokens, perhaps in proportion to their activation, may offer a better representation of the input.

We described how neuron embeddings can be used to measure neuron polysemanticity, which could be very useful for better evaluating SAEs. We also provided a proof-of-concept demonstrating how we can integrate information from neuron embeddings into the SAE loss. Applying this to a toy MLP model trained on MNIST showed several interesting effects, appearing to trade-off decreased reconstruction accuracy and activation sparsity for increased monosemanticity, as well as significantly decreasing the proportion of dead neurons.

We note that this is early-stage research on a small toy model, so it remains unclear how these results would transfer to larger models. Applying neuron embeddings as an evaluation metric for SAEs trained on real-world language models, as well as experimenting with the neuron embedding loss when training such SAEs, would both be very interesting directions for future work.

Impact Statement

This work presents a new method with applications in mechanistic interpretability of vision and language models. There are no specific ethical implications or societal consequences of this work that we feel need to be highlighted here.