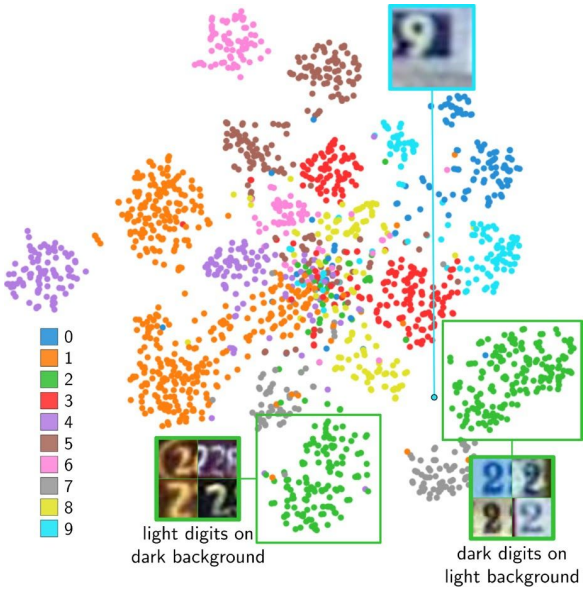
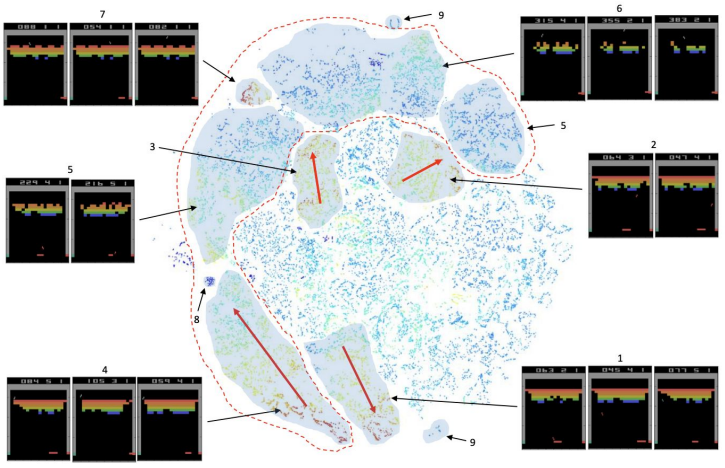
Interpreting neural networks using neural dimensionality reduction

Neural networks are complex predictive models that are difficult to understand. A recent approach to get more insight into the behavior and choices of a neural network are neural dimensionality reduction techniques. In these approaches, the high dimensional neural activations are projected down to two dimensions such that they can be visualized as a scatterplot. Neurons with similar activations are projected close to each other.





Your task is to train a neural network of choice (either a network from a previous project or course, a pre-existing one, or a network you build during the project on an interesting dataset). In case you have no preference, we recommend sticking to image classification (MNIST, CIFAR, ImageNet, COCO datasets, in order of difficulty), or sentiment analysis (twitter dataset, or a certain book, with word embeddings, BERT or GPT-3 network).
Next, you try to uncover insights into how the neural network makes predictions. Are there any clusters of interest? What does the cluster mean? Can you spot mistakes by the model or in the dataset? Your task is to design multiple visualizations and link them together through interaction to help experts understand the neural network.

Relevant paper: [PDF] ieee.org (Visualizing the hidden activity of artificial neural networks)
[PDF] jmlr.org (Graying the black box: understanding DQNs)