Nicole Meister ELE'22

t-SNE Clustering on Chest X-Ray Images for better COVID identification

Introduction

- Half of COVID-19 patients have pneumonia-like thickening of the lungs, while other half have clearer lungs
- Machine learning can assist in rapid classification of these chest X-Ray images as COVID-19, normal, or pneumonia
- Clustering can create groups based on initially hidden features

Problem

- t-SNE[1] does not preserve the embedding to cluster new images the visualization has not seen before
- High dimensionality reduction can make it difficult capture all the information and with t-SNE clustering, it is often difficult to understand what the machine is learning

1. van der Maaten, L. & Hinton, G. (2008). Visualizing Data using t-SNE . *Journal of Machine Learning Research*, 9, 2579--2605.

Methods

- Open t-SNE[1]: does not preserve the embedding to cluster new images the visualization has not seen before
- Silhouette Score to identify mis-classified images
- Sobel Filter: Edge detection filter
- Tree-SNE[2]:

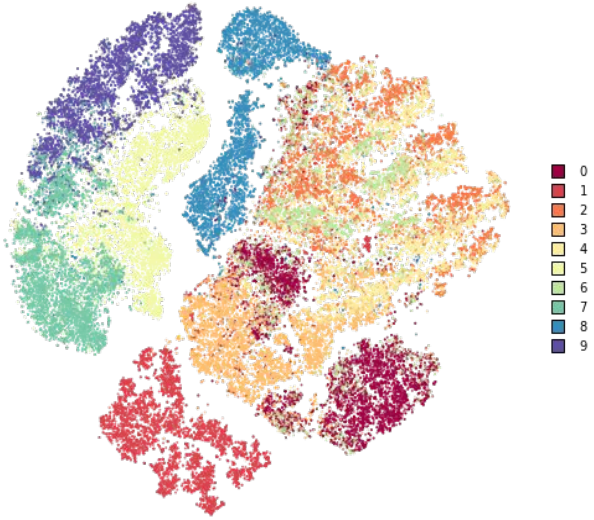
1. Pavlin G. Poličar, Martin Stražar, Blaž Zupan .openTSNE: a modular Python library for t-SNE dimensionality reduction and embedding. bioRxiv 731877
2. Robinson, Isaac and Emma Pierce-Hoffman. "Tree-SNE: Hierarchical Clustering and Visualization Using t-SNE." ArXiv abs/2002.05687 (2020):

Open t-SNE

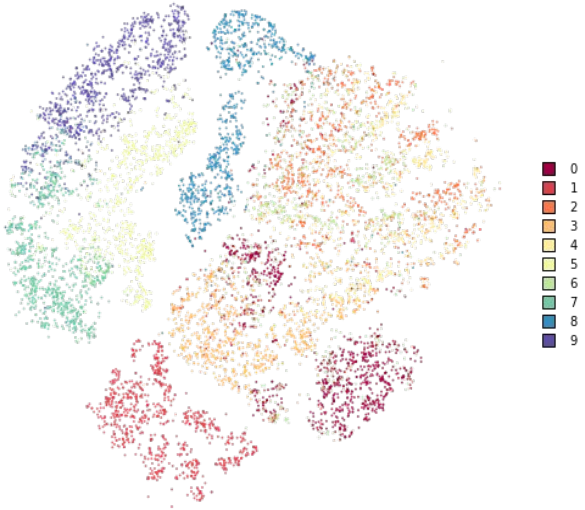
- Ability to add new data points to existing embeddings
  - Add new point to the graph in the middle, see where the KL divergence is lowest for that point
- Explore variations of this method (Early Exaggeration, Optimization) to improve 1-KNN accuracy

Pavlin G. Poličar, Martin Stražar, Blaž Zupan .openTSNE: a modular Python library for t-SNE dimensionality reduction and embedding. bioRxiv 731877

Open t-SNE: Fashion MNIST

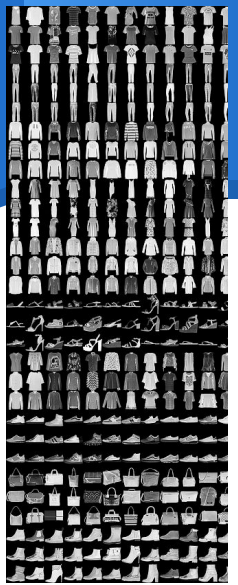


60000 Points for Training



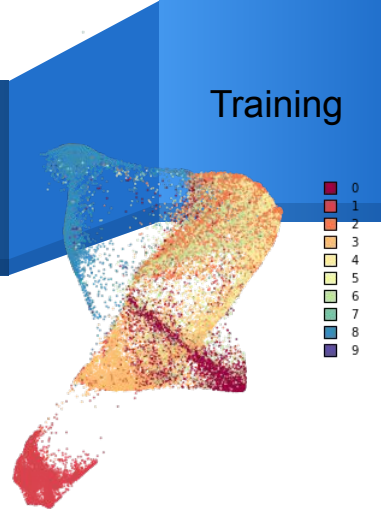
10000 Points for Testing

1-KNN: 0.79



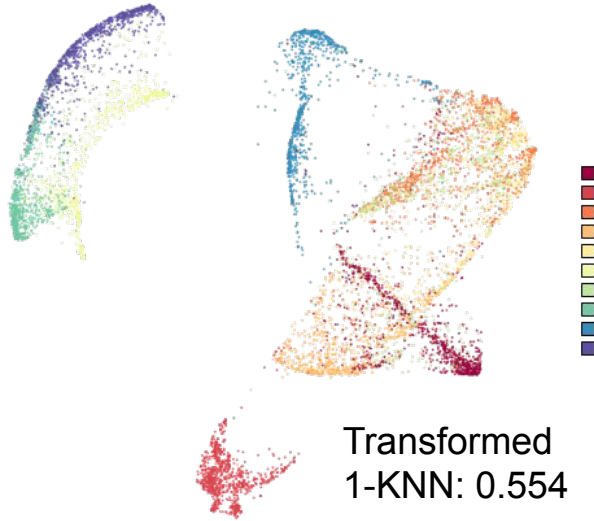
Label	Description
0	T-shirt/top
1	Trouser
2	Pullover
3	Dress
4	Coat
5	Sandal
6	Shirt
7	Sneaker
8	Bag
9	Ankle boot

Open t-SNE: Advanced Usage – Early Exaggeration

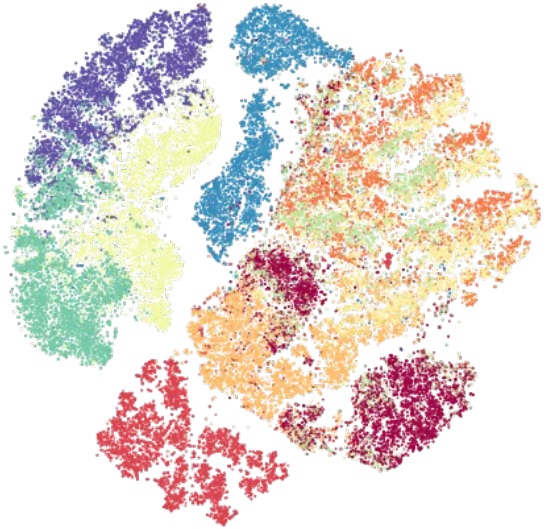


Early Exaggeration: all pairwise probabilities relating to map points (p) are multiplied by a factor so that the q probabilities are too small to model the p.

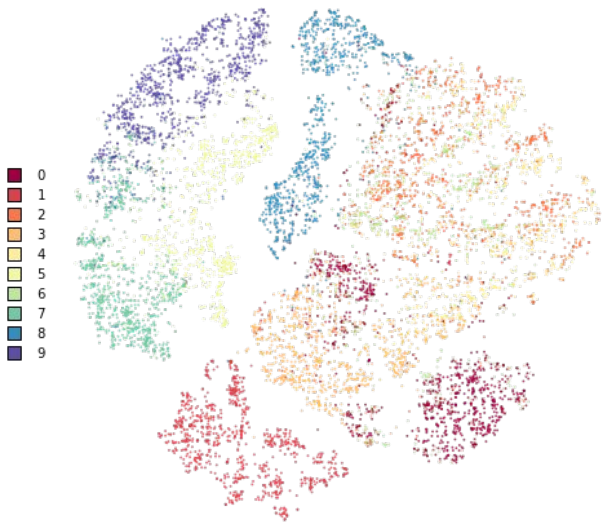
The optimization is encouraged to model the large p with large qs ⟶ tight widely separated clusters and thus easier for cluster to move around



Open t-SNE: Advanced Usage - Optimization



Optimization (Momentum = 0.8)

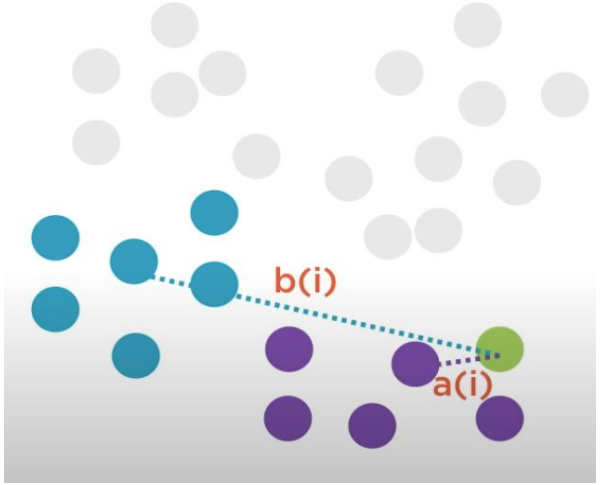


Transformed (1-KNN: 0.7682)

Optimized transform (1-KNN: 0.8227)

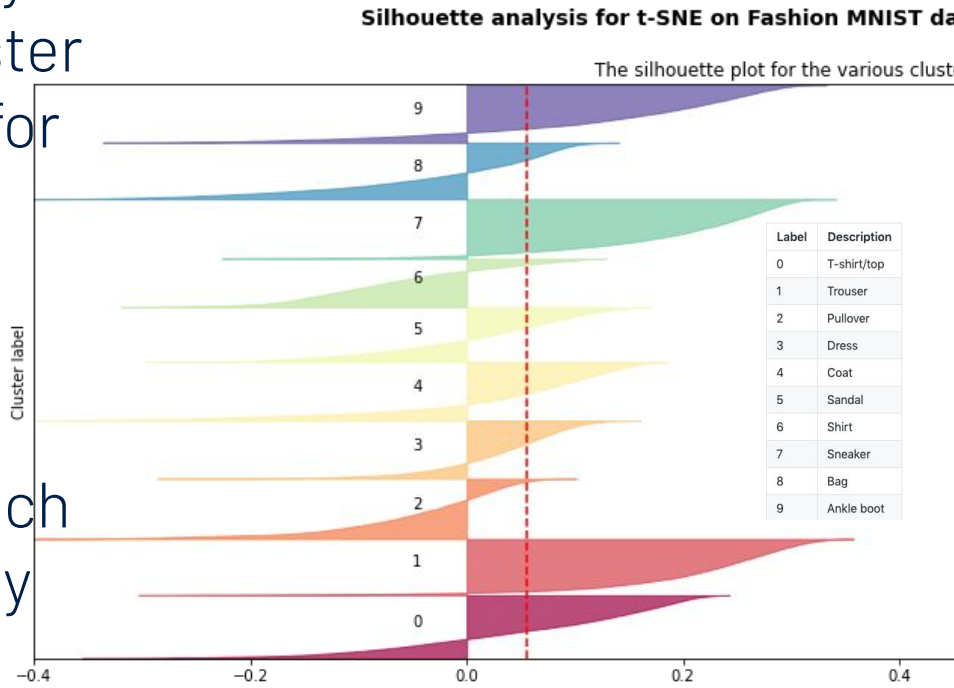
Silhouette Score: Concept

- Silhouette Score: calculated using the mean intra-cluster distance (`a`) and the mean nearest-cluster distance (`b`) for each sample.
- The Silhouette Coefficient for a sample is `(b - a) / max(a, b)`.
  - 1: best
  - -1: worst
  - 0: indicative of overlapping clusters



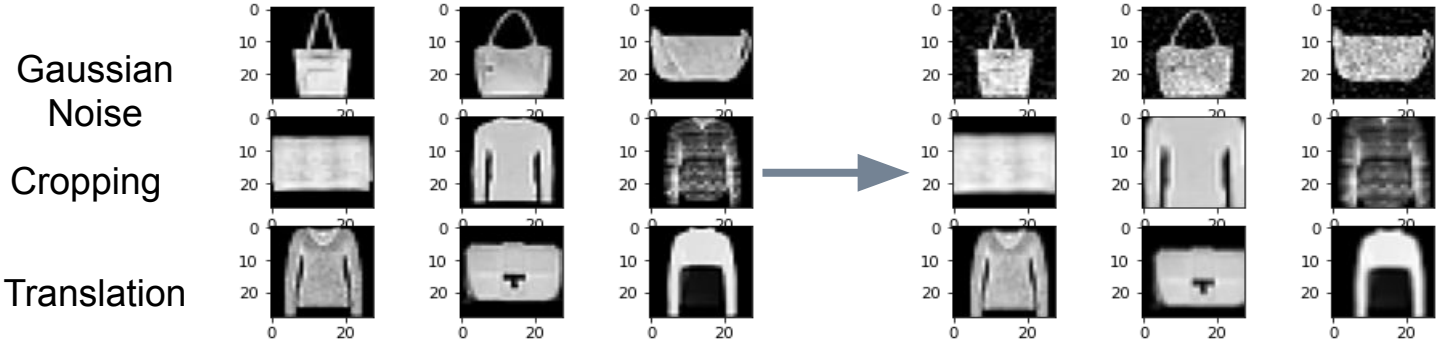
Silhouette Score: Method

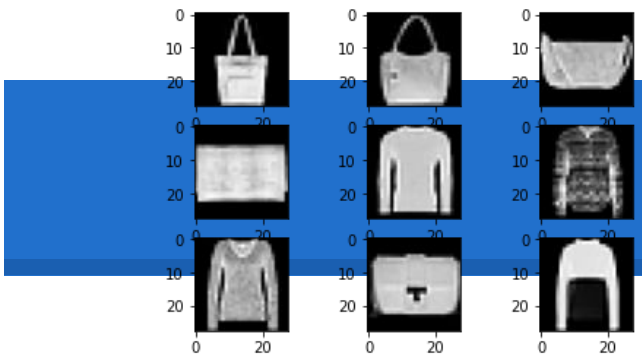
- 10-KNN Classifier to identify class of each point in a cluster
- Calculate silhouette score for each point in a cluster
  - Understand how well matched point is to its own cluster
- Plot silhouette scores of each point in a cluster grouped by each class

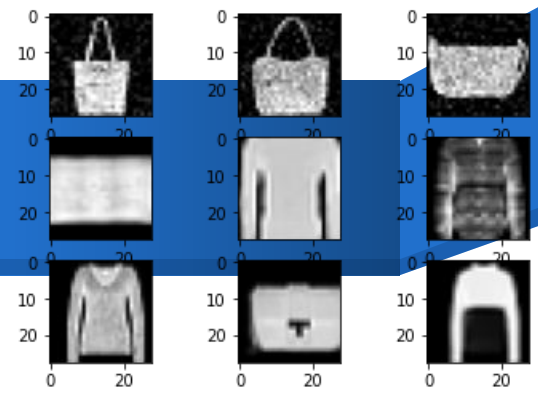


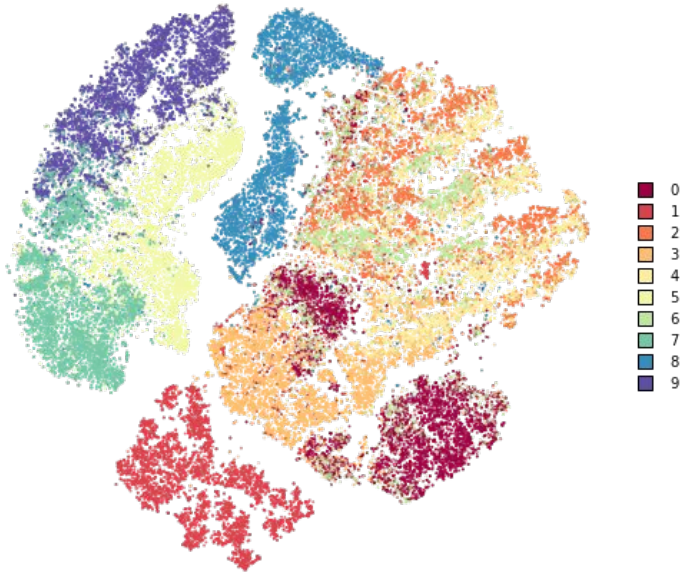
Using SS Scores to inform another iteration of clustering

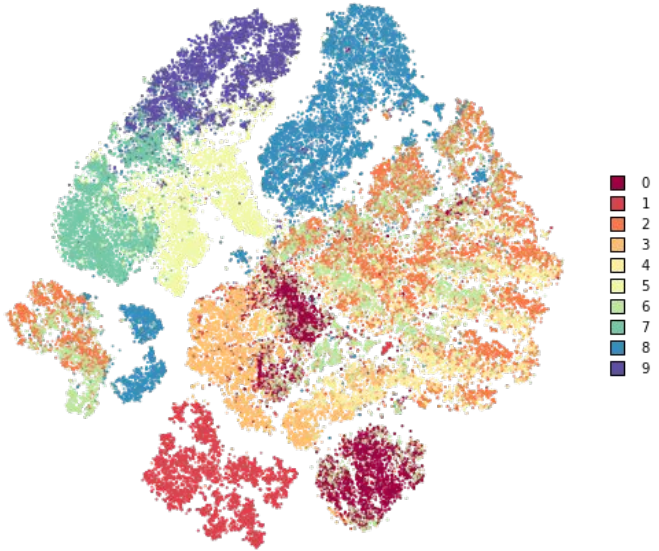
- Data Augmentation: Apply random levels of Gaussian Noise, Cropping, Translation to the classes (2, 6, 8)
- Increase training size from 60k to 78k images (18k are repeat images with these affine transforms applied)



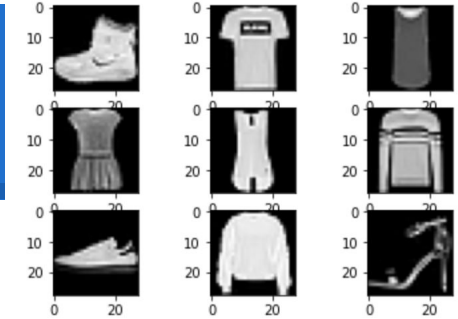


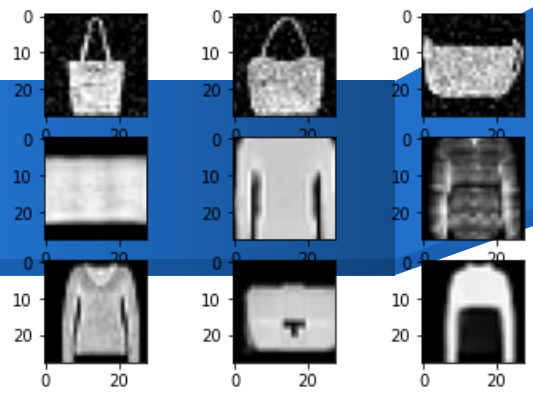


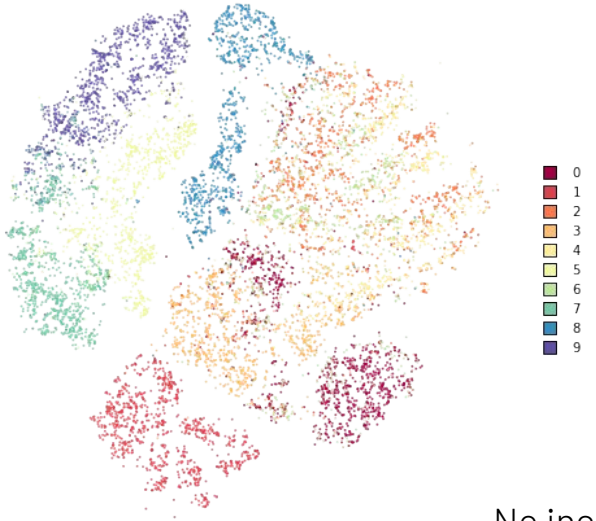

0
1
2
3
4
5
6
7
8
9

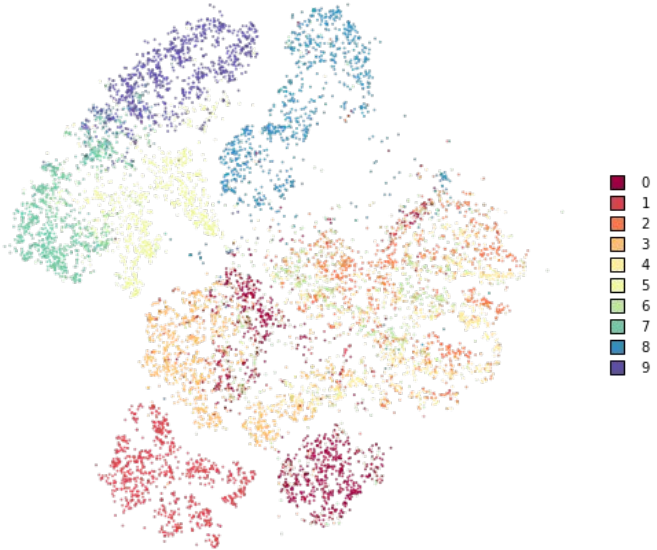

0
1
2
3
4
5
6
7
8
9

```
Train: X=(60000, 28, 28), y=(60000,)
Test: X=(10000, 28, 28), y=(10000,)
```







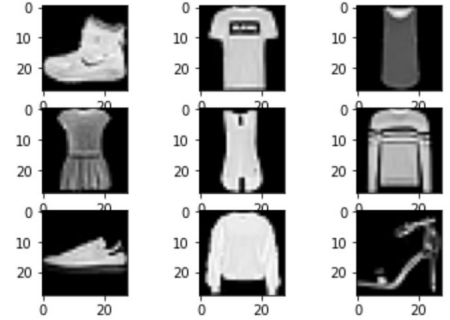


Accuracy: 0.80

No increase in accuracy

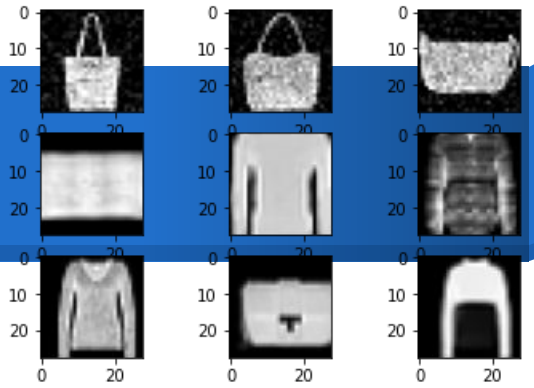
Accuracy: 0.78

```
Train: X=(60000, 28, 28), y=(60000,)
Test: X=(10000, 28, 28), y=(10000,)
```



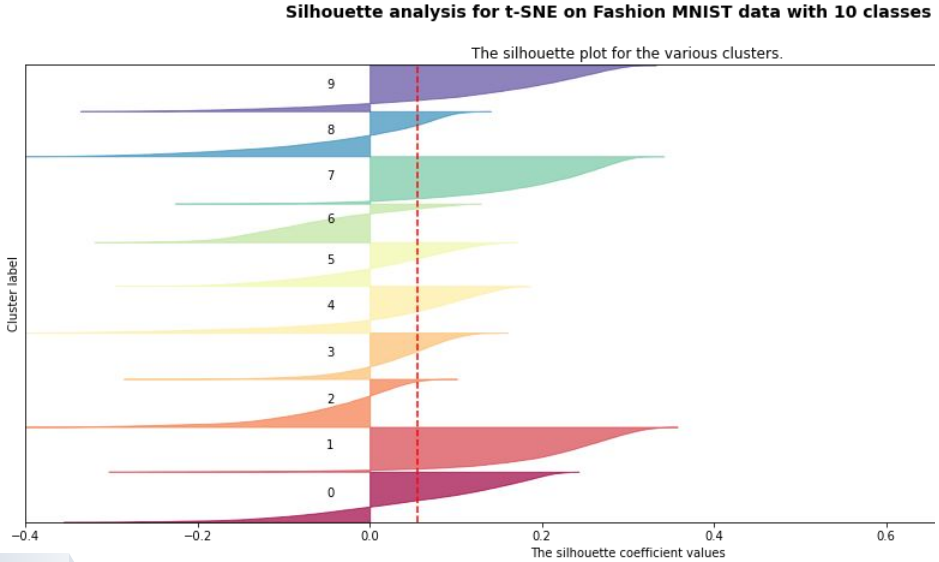
```
0's avg ss: 0.059
1's avg ss: 0.21
2's avg ss: -0.046
3's avg ss: 0.030
4's avg ss: 0.033
5's avg ss: 0.0093
6's avg ss: -0.066
7's avg ss: 0.19
8's avg ss: -0.024
9's avg ss: 0.14
```

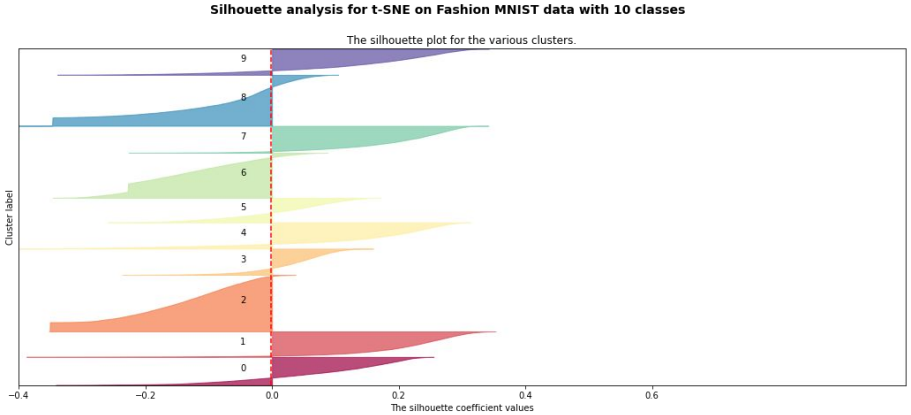
0.05353



```
0's avg ss: 0.072
1's avg ss: 0.21
2's avg ss: -0.16
3's avg ss: 0.031
4's avg ss: 0.13
5's avg ss: 0.011
6's avg ss: -0.14
7's avg ss: 0.191
8's avg ss: -0.105
9's avg ss: 0.141
```

-0.0016

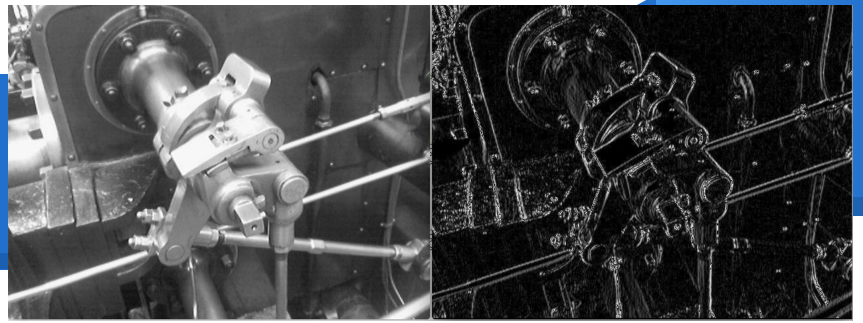




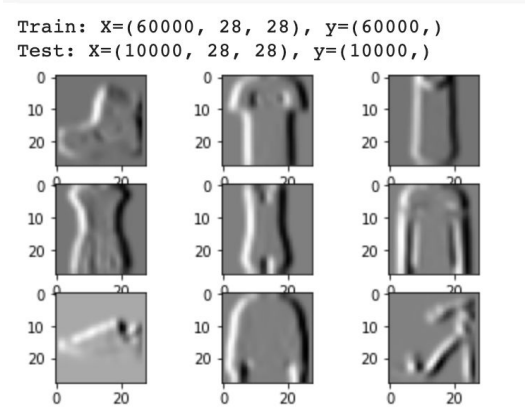
Results with Silhouette Score

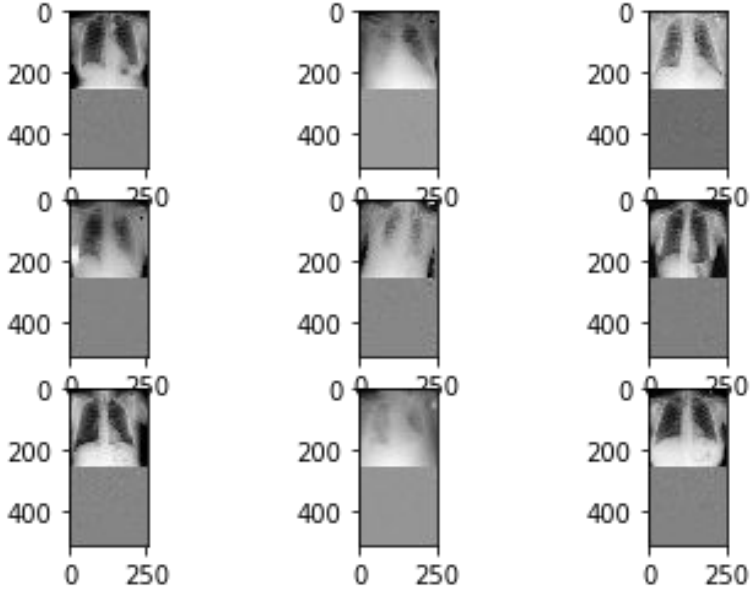
Augmenting the dataset with labels that had low silhouette scores actually decrease the overall silhouette score

Sobel Filter



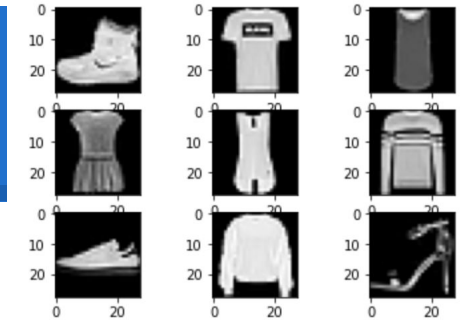
- Apply an edge detection filter to identify the outline of object, ribs, etc
- Use edge information to aid clustering



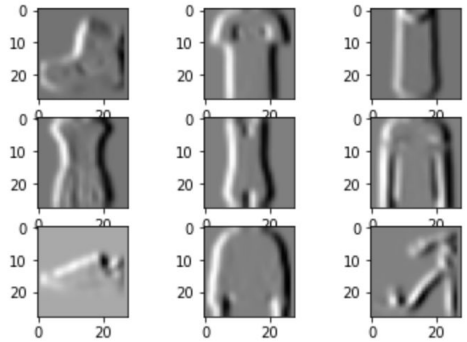


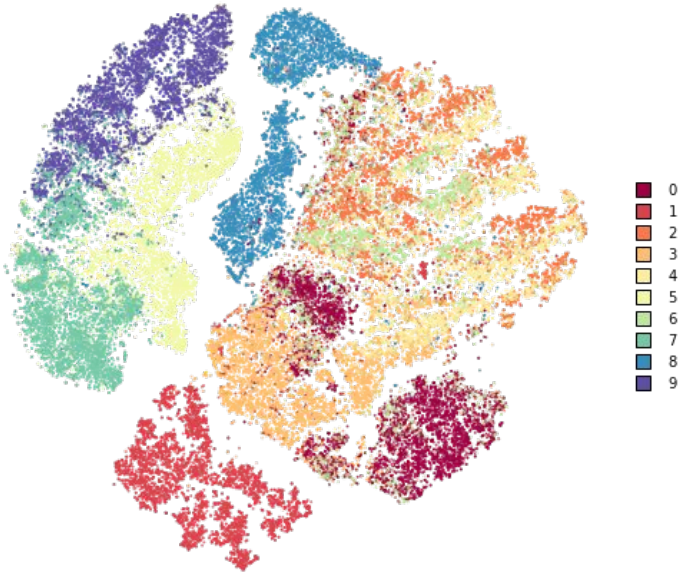
Fashion MNIST vs Sobel X Filter

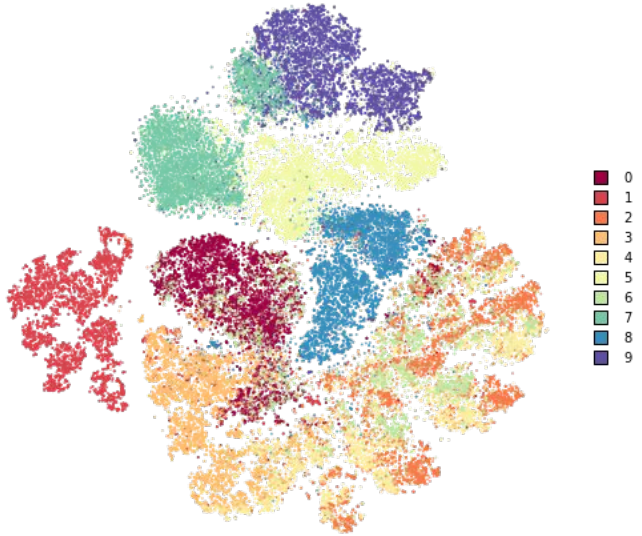
```
Train: X=(60000, 28, 28), y=(60000,)
Test: X=(10000, 28, 28), y=(10000,)
```

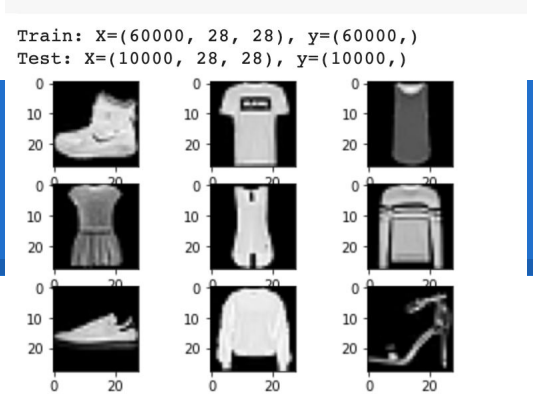


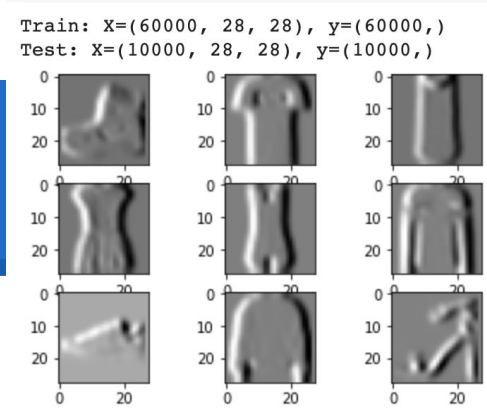
```
Train: X=(60000, 28, 28), y=(60000,)
Test: X=(10000, 28, 28), y=(10000,)
```

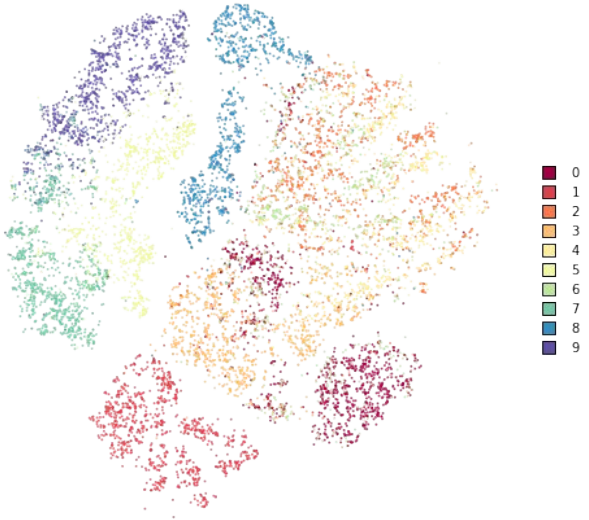


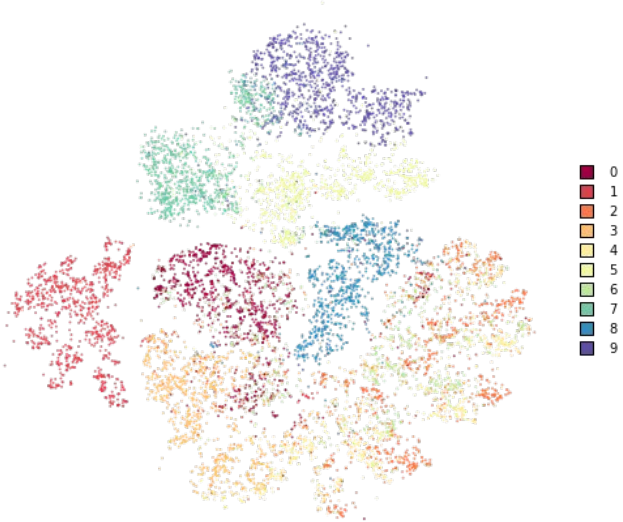






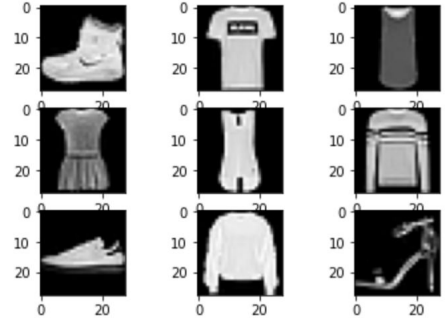






Similar accuracy: 0.8

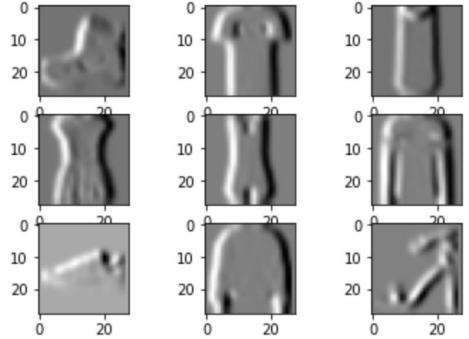
Train: X=(60000, 28, 28), y=(60000,)
Test: X=(10000, 28, 28), y=(10000,)



```
0's avg ss: 0.059
1's avg ss: 0.21
2's avg ss: -0.046
3's avg ss: 0.030
4's avg ss: 0.033
5's avg ss: 0.0093
6's avg ss: -0.066
7's avg ss: 0.19
8's avg ss: -0.024
9's avg ss: 0.14
```

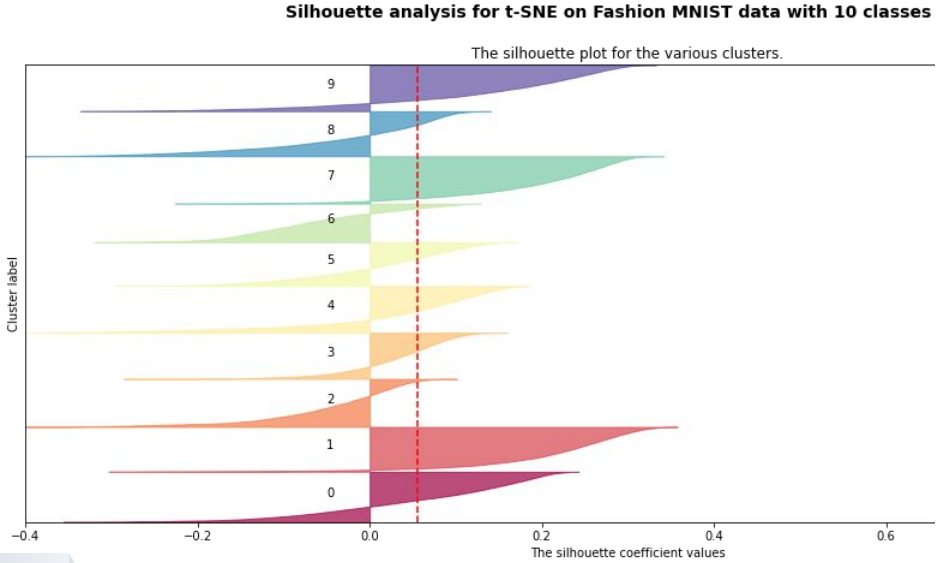
0.05353

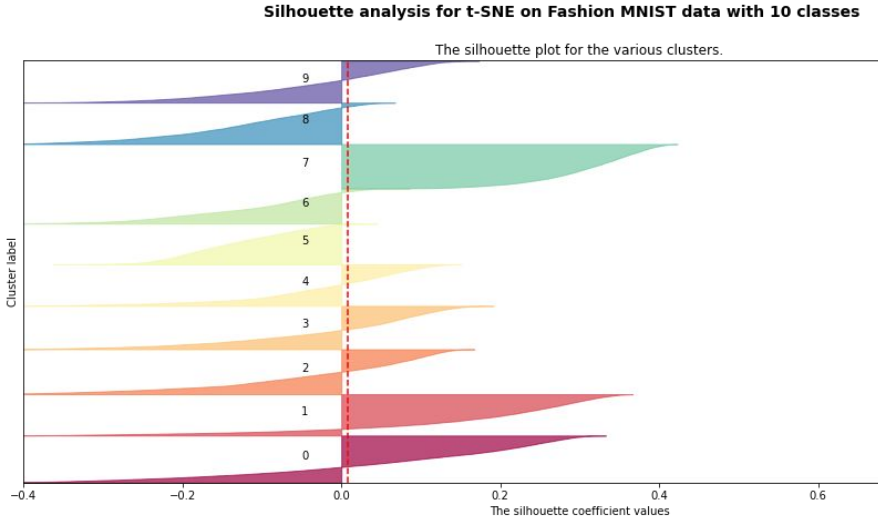
Train: X=(60000, 28, 28), y=(60000,)
Test: X=(10000, 28, 28), y=(10000,)



```
0's avg ss: 0.063
1's avg ss: 0.15
2's avg ss: -0.032
3's avg ss: -0.014
4's avg ss: -0.025
5's avg ss: -0.138
6's avg ss: -0.12
7's avg ss: 0.31
8's avg ss: -0.13
9's avg ss: -0.035
```

0.00289

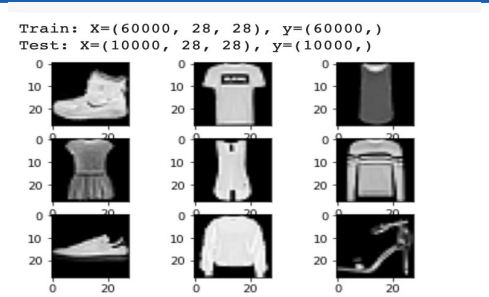


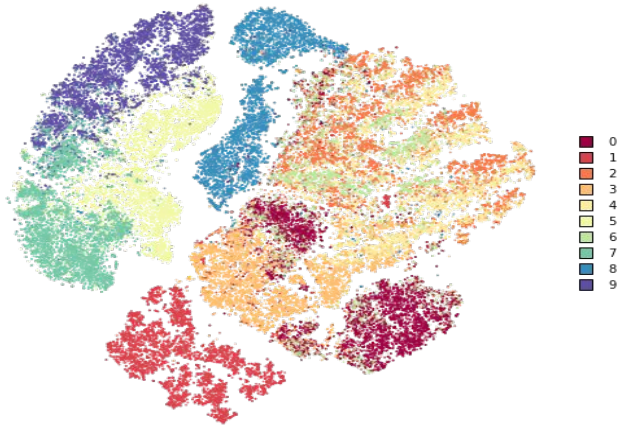


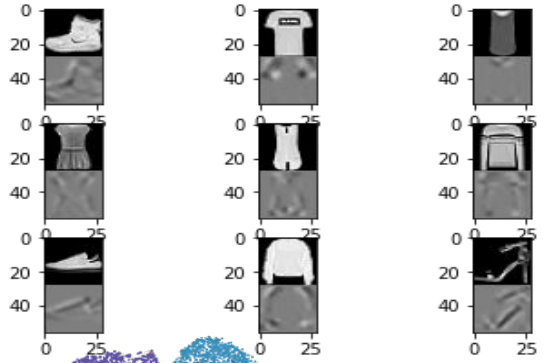
Results on applying Sobel Filter to Fashion MNIST

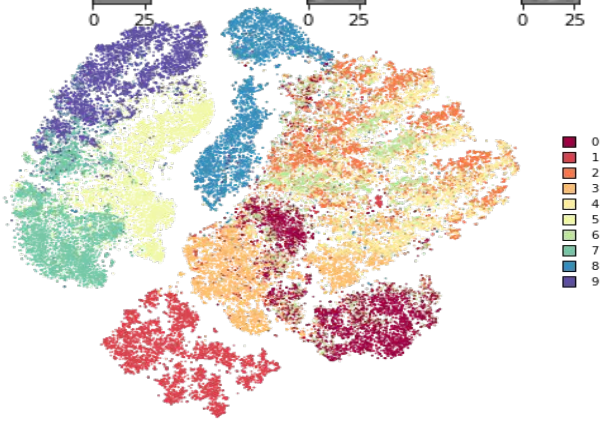
Comparing regular image vs image with Sobel filter, no increase in Silhouette score and similar 1-KNN accuracy

Concatenating Sobel + Fashion MNIST

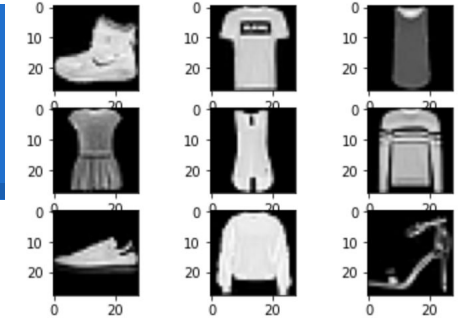


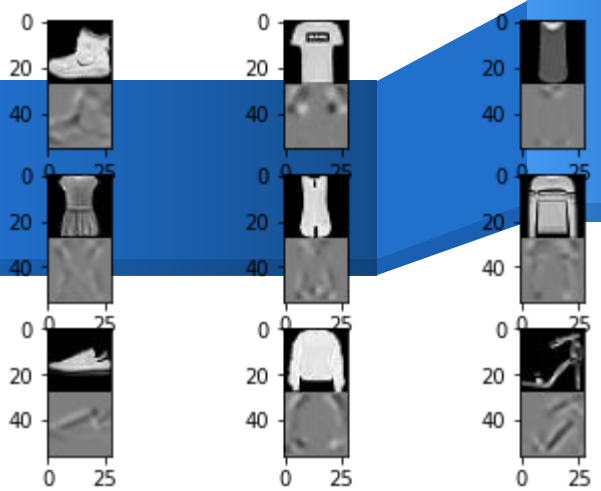


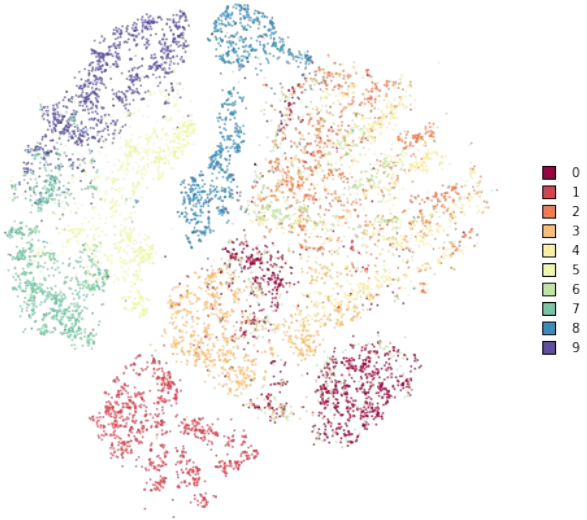


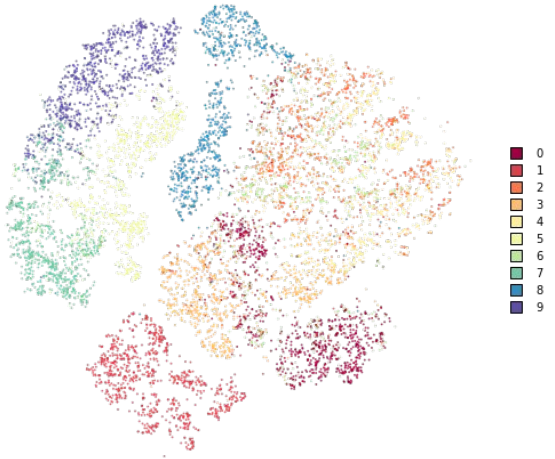


```
Train: X=(60000, 28, 28), y=(60000,)
Test: X=(10000, 28, 28), y=(10000,)
```



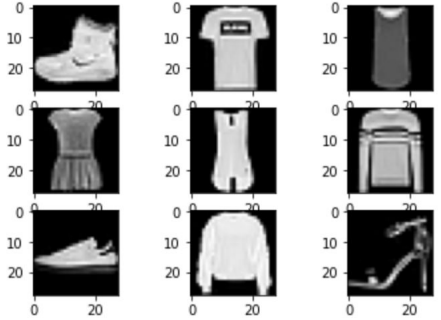






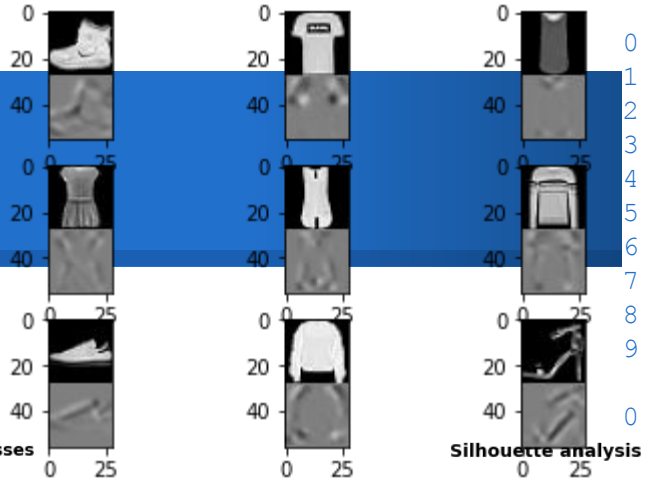
Similar accuracy: 0.8 vs 0.78

```
Train: X=(60000, 28, 28), y=(60000,)
Test: X=(10000, 28, 28), y=(10000,)
```



```
0's avg ss: 0.059
1's avg ss: 0.21
2's avg ss: -0.046
3's avg ss: 0.030
4's avg ss: 0.033
5's avg ss: 0.0093
6's avg ss: -0.066
7's avg ss: 0.19
8's avg ss: -0.024
9's avg ss: 0.14
```

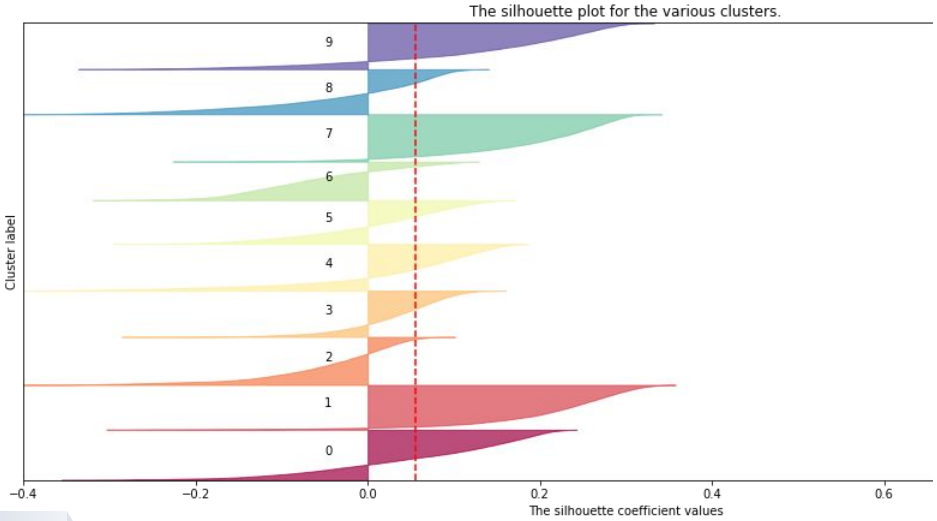
0.05353



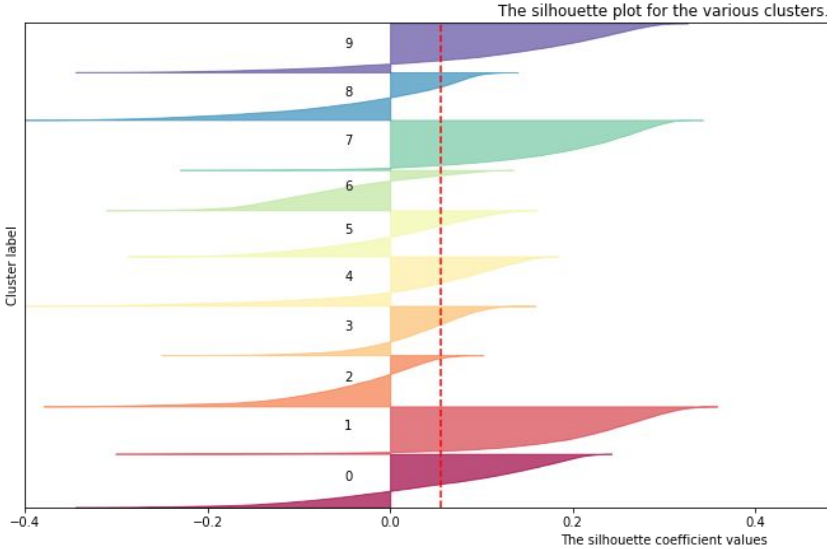
```
0's avg ss: 0.057787619
1's avg ss: 0.213689674
2's avg ss: -0.04236055
3's avg ss: 0.027889015
4's avg ss: 0.032131915
5's avg ss: 0.003152680
6's avg ss: -0.05998175
7's avg ss: 0.195011024
8's avg ss: -0.02314552
9's avg ss: 0.131897352
```

0.0556

Silhouette analysis for t-SNE on Fashion MNIST data with 10 classes



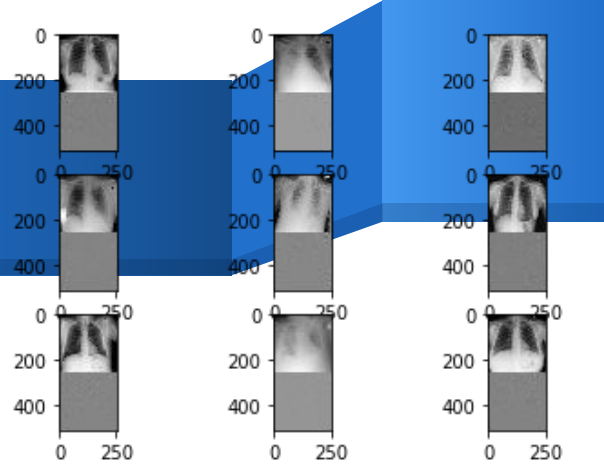
Silhouette analysis for t-SNE on Fashion MNIST data

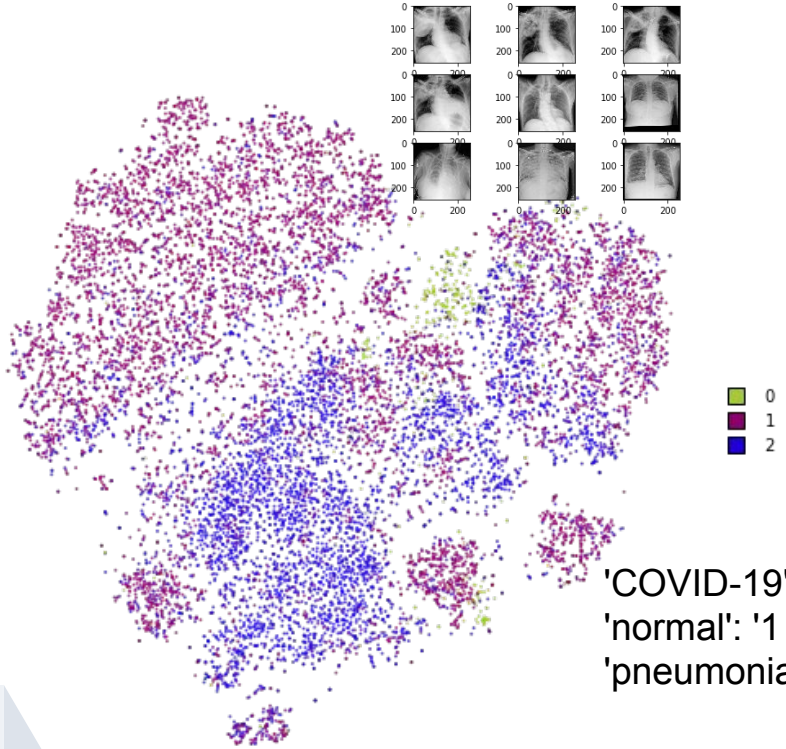


Result of concatenating Sobel + Fashion MNIST

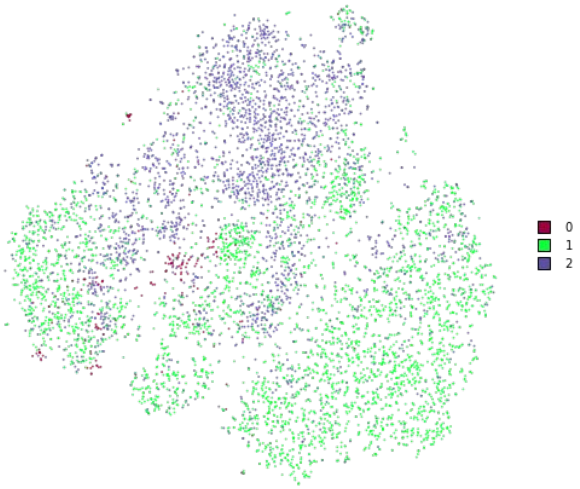
Even after apply different intensities of the sobel filter (25%, 50%, 75%, 100%), the Silhouette Score, increased by 0.002 and the 1-KNN accuracy stayed the same

Sobel on Chest X-Ray Images





'COVID-19': 0
'normal': '1
'pneumonia': 2



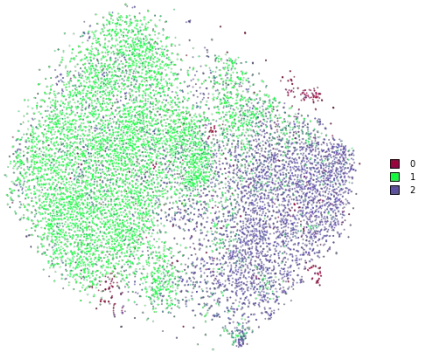
Results from Sobel on Chest X-Ray Images

- No distinct clusters separating the labels: COVID-19, pneumonia, normal
- Low 1-KNN accuracy

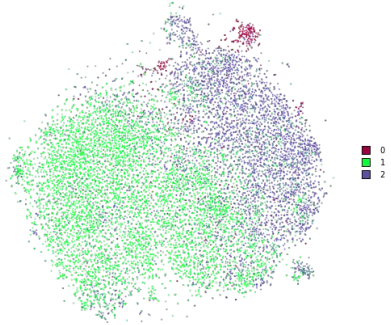
Clustering on Features from DenseNet

Objective: Obtain a feature map from DenseNet and feed feature vector into t-SNE clustering

Result: Still no distinct clusters, many overlapping clusters



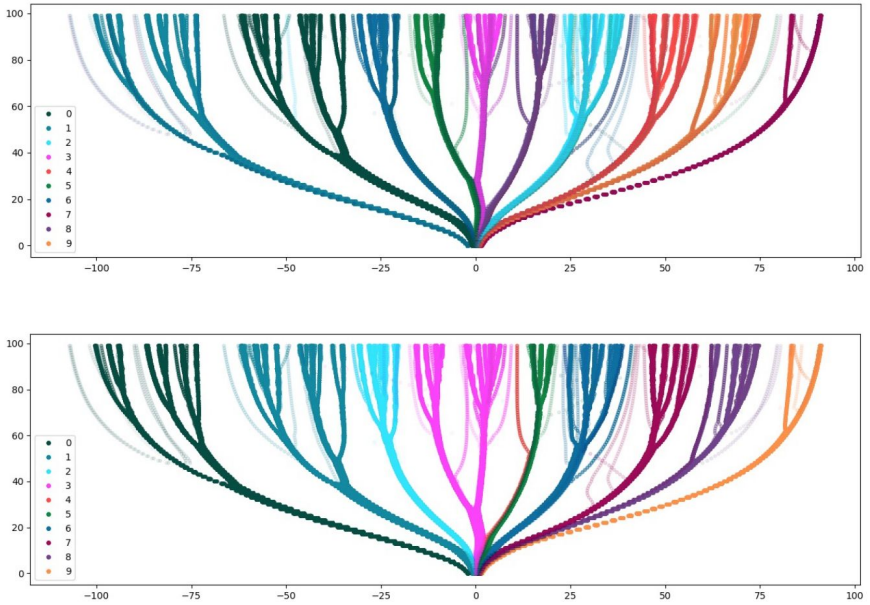
Clustering on features from COVID chest X-Rays



Clustering on features from Sobel filtered COVID chest X-Rays

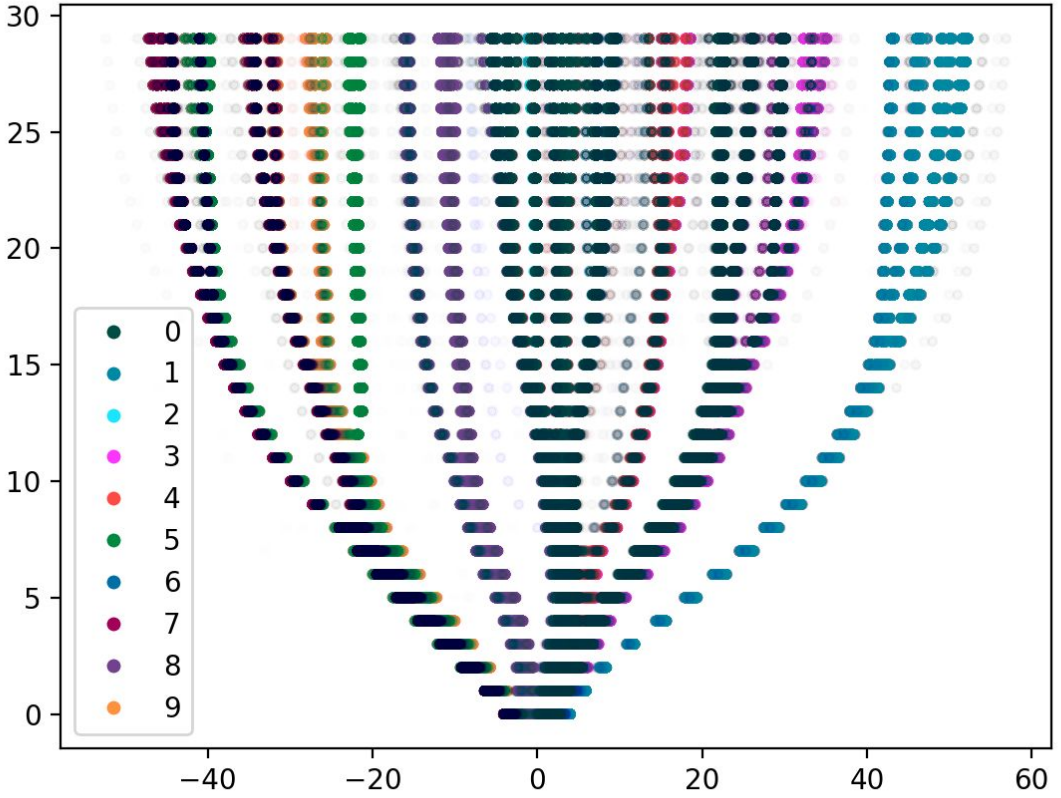
Tree-SNE

creating t-SNE embeddings with increasingly heavy tails to reveal increasingly fine-grained structure, and then stacking these embeddings to create a tree-like structure



Figures from tree-SNE paper

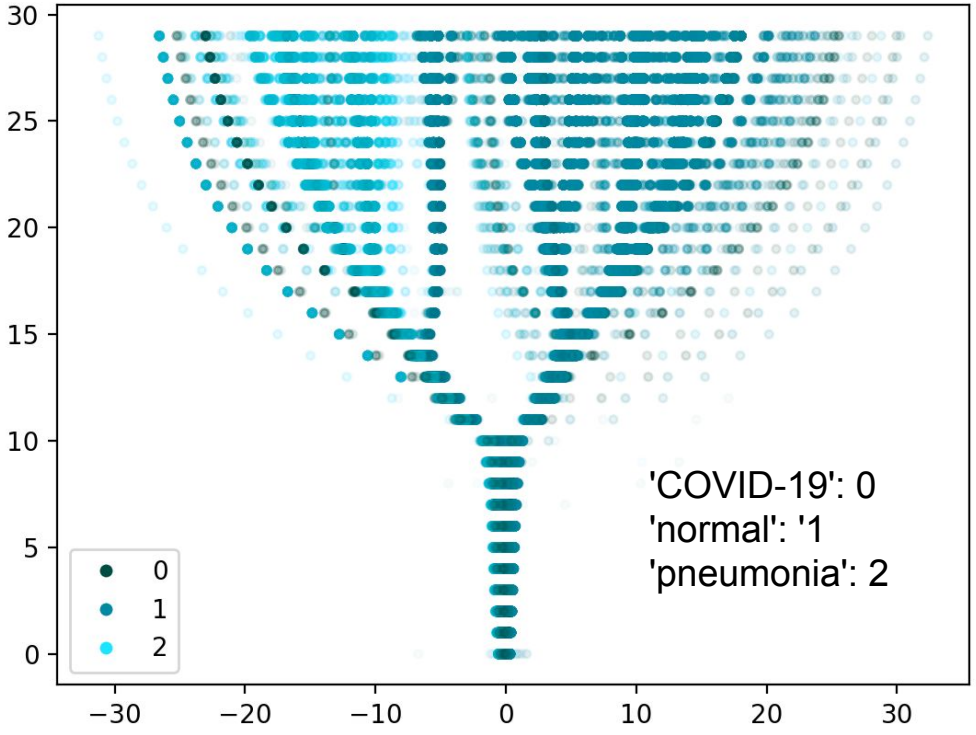
Tree-SNE on Fashion MNIST



Label	Description
0	T-shirt/top
1	Trouser
2	Pullover
3	Dress
4	Coat
5	Sandal
6	Shirt
7	Sneaker
8	Bag
9	Ankle boot

Tree-SNE on Chest X-Ray Features from Dense-Net

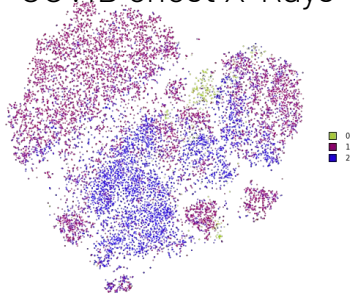
1024 features produced by Densenet



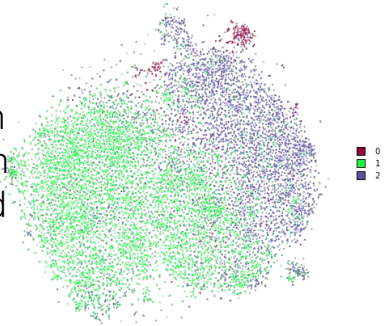
Conclusion

- Open t-SNE is useful for adding new data points to an existing embedding.
- The attempts to improve visualization classification by targeting the silhouette score and concatenating sobel filters were not effective in improving the 1-KNN accuracy of the t-SNE clustering.
- Most effective method of improving visualization was the clustering on features extracted from the Sobel filter COVID chest X-Rays although extremely distinct clusters still did not form

Clustering on features from Sobel filtered COVID chest X-Rays



Clustering on features from Sobel filtered COVID chest X-Rays



Future Work

- Ensemble clustering to combine t-SNE improvement methods to produce more accurate classifications
- Applying early exaggeration or other optimization methods to improve the clustering of features extracted from the Sobel filtered chest x-ray images.
-