Summary

The aim of this research was to apply tools from Topological Data Analysis to the CIFAR-10 dataset. I selected two subsets of the dataset, for which I computed their death vectors and persistence landscape. The results reveal that these tools can provide relevant information to for image classification problems.

Dataset

The two subsets that I used were the cats and trucks image sets from CIFAR-10. The training set consisted of 5000 images per class, and the test set of 1000 images per class, all with a with a resolution of 32x32.

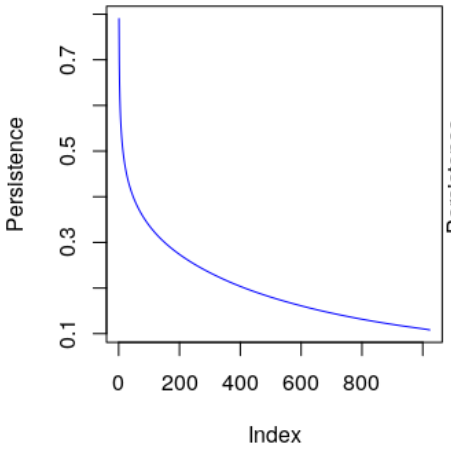
Sample cat image

Sample truck image

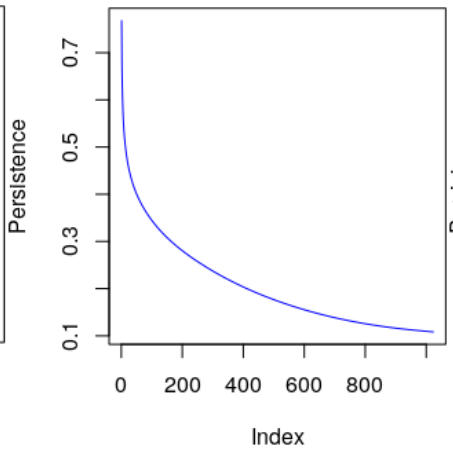
I encoded each image as a list of 1024 points in $\mathbb{R}^5$, with each point representing the Red, Green, Blue color values and X, Y coordinates of a pixel. I used the Euclidean distance.

Computing death vector and persistence landscape

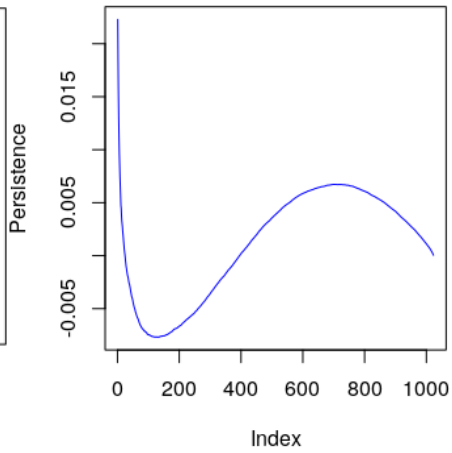
Using the TDA pipeline from Dr. Bubenik's research group I computed the death vector and persistence landscape for each image, as well as the average death vector and average persistence landscape and the difference between these.



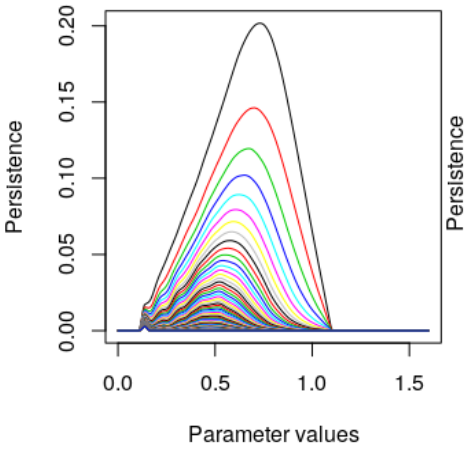
Cats average death vector



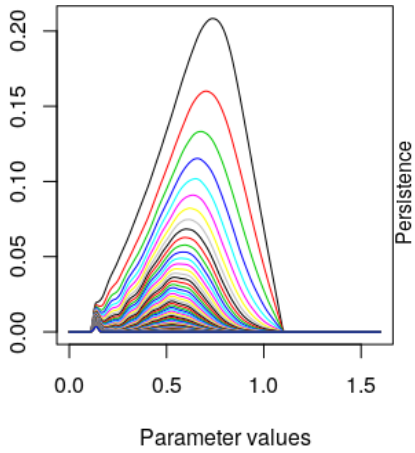
Trucks average death vector



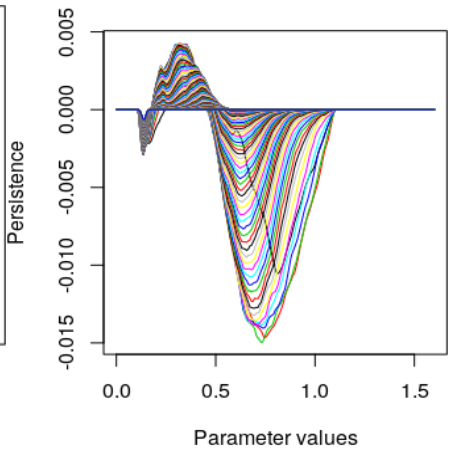
Average death vector difference



Cats average persistence landscape



Trucks average persistence landscape



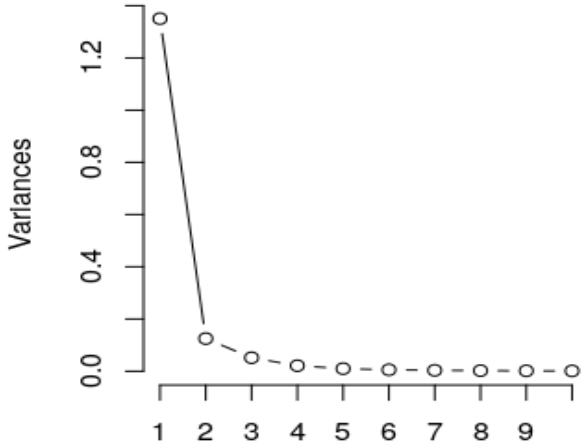
Average persistence landscape difference

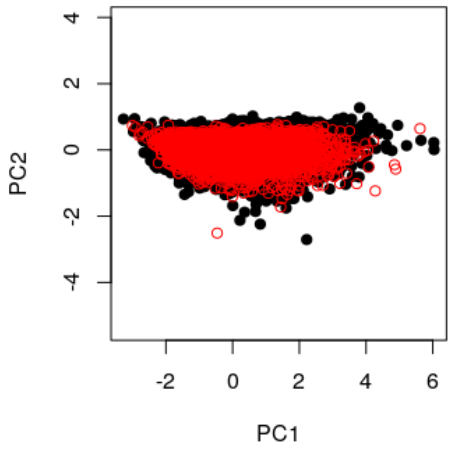
Permutation test

On the death vectors I performed the permutation test with 10,000 permutations. On the persistence landscapes I performed the permutation test with 1,000 permutations. On both cases the p value obtained was of 0.

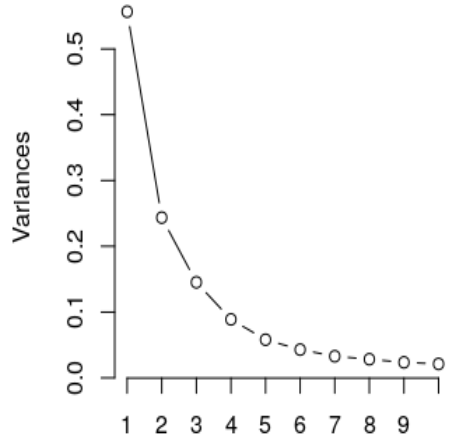
Principal component analysis

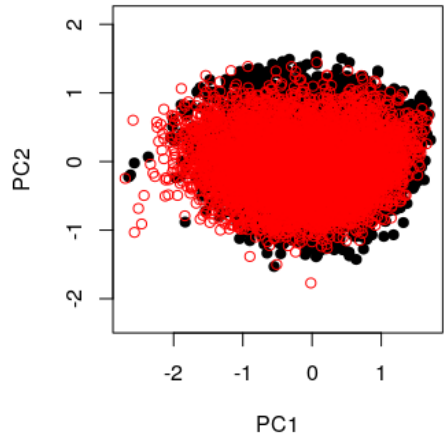
On H_0:





On H_1:





The plots show a total overlap between both subsets.

Support vector machine and neural network

I created two support vector machines that classified the images using either their death vectors or the persistence landscape. Additionally, I trained a deep feedforward neural network using Tensorflow that received both the death vector and the persistence landscape as inputs.

The structure of the neural network was as follows:

- input layer with 21,123 nodes (1023 for death vector and 201x100 for persistence landscape)
- 3 hidden layers with 10,000 nodes
- output layer with one node (binary classifier)
- optimizer: Adam
- activation function: ReLU

Note: the number of hidden layers and nodes per hidden layers set to these values due to hardware limitations (this is the biggest structure that would fit in my GPU's memory). Using more hidden layers with more nodes would likely have given a better accuracy.

I tested all three models an the test set and measured their accuracy on the individual classes as well as the entire test set. Here are the results:

Model	Cats	Trucks	Cats & Trucks
SVM (H_0)	72.2%	68.0%	70.1%
SVM (H_1)	70.5%	68.3%	69.4%
NN (H_0 & H_1)	73.9%	72.6%	73.25%

Conclusion

These accuracy values reveal that the death vectors and persistence landscape contain relevant information that could help more complex models improve their performance (see Improved Image Classification using Topological Persistence).