Topological Data Analysis of Convolutional Neural Networks' Weights on Images

Rickard Brüel Gabrielsson
Stanford University Department of Computer Science

Abstract

The topological properties of images have been studied for a variety of applications, such as classification, segmentation, and compression. In the application of image classification, high classification accuracy has been achieved by machine learning models, especially **convolutional neural networks**. In our project, we apply topological data analysis to describe, visualize, and analyze topological properties of the weights learned by a CNN classifier trained on digit images from the MNIST data set.

Introduction

- **Image Classification:** Convolutional neural networks (CNNs), with tremendous success, assign weights to small regions (filters) of the pixels of an image
- **Problem:** However, how the high-level features learned by CNNs contribute to their success is not fully understood
- **Related Work:** Topology data analysis (TDA) has been applied by Carlsson [1] and Lee et al [3] to study natural image statistics and to generate dimensionality-reduced topological networks from data, since natural images contain rich structures within a high-dimensional point cloud where topological properties are far from obvious. Lee et al. observed that high-contrast 3×3 optical patches in natural images were concentrated around a non-linear continuous 2-dimensional submanifold resembling blurred step edges.
- **Project Goal:** Gain insight into what features CNNs learn from small image patches and how the CNN weights evolve over training by generating and analyzing networks using Carlsson's Mapper Method on the weights of CNNs image classifiers
- **Significance:** By applying network analysis to the **new context** of graphs generated by TDA on the learned weights of a CNN image classifier, we enrich network analysis as well as the fields of TDA and machine learning

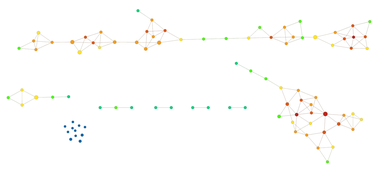
Methodology

- **Dataset Generation:** We trained a multilayer convolutional neural network (CNN) on the MNIST dataset of hand-written digits to obtain the weights of its first convolutional layer, giving us a dataset of 512 weights vectors in $\mathbb{R}^{25}$
- **Network Model:** Used Nearest-Neighbor lenses as reference maps and Variance Normalized Euclidean distance metric along with Ayasdi software. We also tried a model using PCA.
- **Procedure:** MNIST dataset → CNN → First-layer weights → Preprocessing → Mapper Method → Network → Visualization

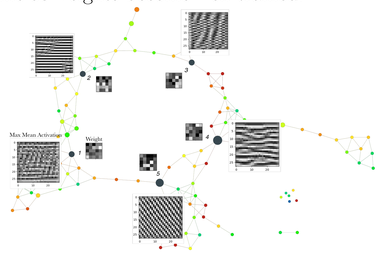
Example MNIST image of size 28×28 pixels in greyscale and label 7

Hypothesis

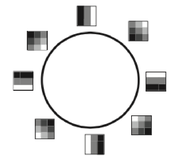
Temporal Network Analysis Hypothesis: Based on the following preliminary results, we hypothesized that the network generated from CNN weights forms into a circular structure over time as weights become well-trained:



A network generated from CNN weights prior to training on MNIST (Gaussian random weight initialization)



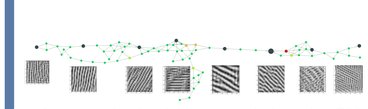
Network after 1 epoch of training (96% training accuracy), along with some of the max mean activations for nodes in the circle



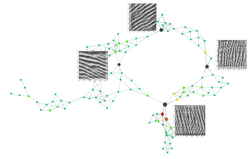
Circle of image patches from by Carlsson et al [2]

Results

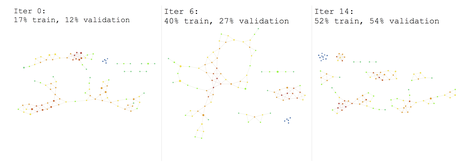
- **Network Structure:** Instead of seeing the weights network converge to a circular structure, we saw circles alternate between formation and breaking apart in the generated network over the course of training the weights (saved every 100 iterations).



Maximally activating inputs at train iteration 700

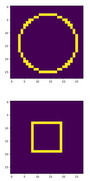


Maximally activating inputs at train iteration 800



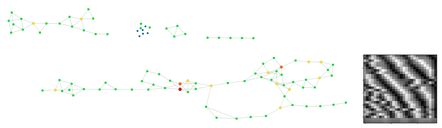
Another sample generated temporal network sequence, but labeled by training accuracy

- **Synthetic Data:** In order to investigate the effects of different image data on the weights learned and the resulting network, we generated and trained the same CNN on synthetic data. We generated two different networks based on the synthetic data (10,000 training data, 2,500 test data, and batch size 200), one after 40 batch iterations and where the network was achieving 100% accuracy, and another after 2,000 batch iterations where the network had achieved a training loss of order 10^{-8}



Synthetic data

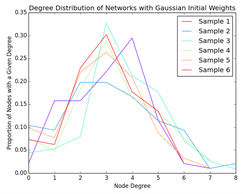
Network and max-activation after 40 batch iterations

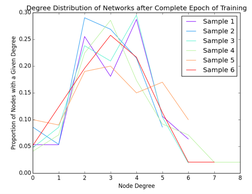


Network and max-activation after 2,000 batch iterations

- **Null Model Comparison:** We noticed that depending on initialization, circular structures may even appear in the random initial weights. Also, networks prior and after 1 epoch of training had similar summary statistics: the sample mean network **diameter** prior to training was 17.8 (standard deviation 2.13) and after 1 epoch of training was 19.5 (stddev 3.62).

Samples of networks generated with Gaussian random weight initialization





Mathematical Background

- **Density Filtrations:** We first filter dataset X to a core subset $X(k,p)$ which is the p percent of points $x \in X$ with smallest distance between x and its k-th nearest neighbor
- **Mapper Method:** Gives a simplicial complex (a network) from which we can discern qualitative properties of X:
  1. Define a reference map $f : X \to Z$, where X is the given point cloud and Z is the reference metric space.
  2. Select a set covering $\mathcal{U}$ of Z.
  3. If $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$, then construct the subsets $X_\alpha = f^{-1}U_\alpha$.
  4. Select a value ϵ as input. Apply the single linkage clustering algorithm with parameter ϵ to the sets X_α to obtain a set of clusters. This gives us a set covering of X parametrized by pairs (α, c), where $\alpha \in A$ and where c is one of the clusters of X_α.
  5. Lastly, construct the simplicial complex whose node set is the set of all possible such pairs (α, c) and where a family $\{(\alpha_0, c_0), \ldots, (\alpha_k, c_k)\}$ spans a k-simplex if and only if the corresponding clusters have a point in common.

Analysis

- We saw interesting properties of the evolution of the CNN weights into a circular structure (for networks constructed from their normalized nearest neighbors). Also, the maximally activating inputs for the weights on the circle resemble the step edges from Carlsson et al [1].
- However, our results don't allow us to reject the null hypothesis. Well-trained weights aren't necessarily significantly circular because initial random weights are occasionally circular. Increase in diameter over training suggests improved expressiveness of weights, while degree distributions suggest optimal weights have Gaussian distribution.
- Training the same CNN on synthetic data supports the original hypothesis, and minimizing the cost function ultimately leads to a circular topology and linear gradient max-activations.
- This is a qualitative study and to get statistically significant quantitative measurements, future work includes training a great number of CNNs on natural images, which proved a bottleneck since it's computationally expensive. This is the first time these methods are applied to CNNs and the first results are promising.

Acknowledgements

This project has shared components with a project in CS224W with Heather Blundell and Dylan Liu, and it is done under the partial supervision of Stanford Professor Gunnar Carlsson, whose prior work is cited under our related work section. Thanks to CS229 teaching staff!

References

[1] G. Carlsson.
Topology and data.
Bulletin of the American Mathematical Society, 46(2):255âĂŞ308, 2009.

[2] G. Carlsson, T. Ishkhanov, V. D. Silva, and A. Zomorodian.
On the local behavior of spaces of natural images.
International Journal of Computer Vision, 76(1):1âĂŞ12, 2007.

[3] A. B. Lee, K. S. Pedersen, and D. Mumford.
The nonlinear statistics of high-contrast patches in natural images.
International Journal of Computer Vision, 54(1):83–103, Aug 2003.