Subspace Clustering Based Analysis of Neural Networks

Uday Singh Saini, Pravallika Devineni, Evangelos E. Papalexakis

MAD Lab @ UCR

Multi Aspect Data

UC RIVERSIDE

Goals of the Paper

- We motivate Sparse Subspace Clustering (SSC) and Centered Kernel Alignment (CKA) as tools to interpret and analyse trained Neural Networks.

SSC : Sparse subspace clustering: Algorithm, theory, and applications. (Elhamifar et al.,2012)
CKA : Similarity of Neural Network Representations Revisited (Kornblith et al., 2019), Algorithms for Learning Kernels Based on Centered Alignment (Cortes et al.,2012)

Background : Sparse Subspace Clustering

- Let $X = [x_1, x_2, \dots, xN] \in \mathbb{R}^{d \times N}$ Represent the Input Embeddings.
- Where each x_i is the latent representation of the i^{th} input example.
- The goal is to learn $C_X = [c_1, c_2, \dots, c_N] \in \mathbb{R}_+^{(N \times N)}$ where entry C_{ij} represents the affinity of x_j to x_i .
- A Naïve formulation of the problem is framed as :

$$\min_{\mathbf{c}_i} ||\mathbf{c}_i||_0 \quad \text{s.t.} \quad \mathbf{x}_i = X\mathbf{c}_i,\ c_{ii} = 0 \quad \forall i \in \{1, \dots, N\}$$

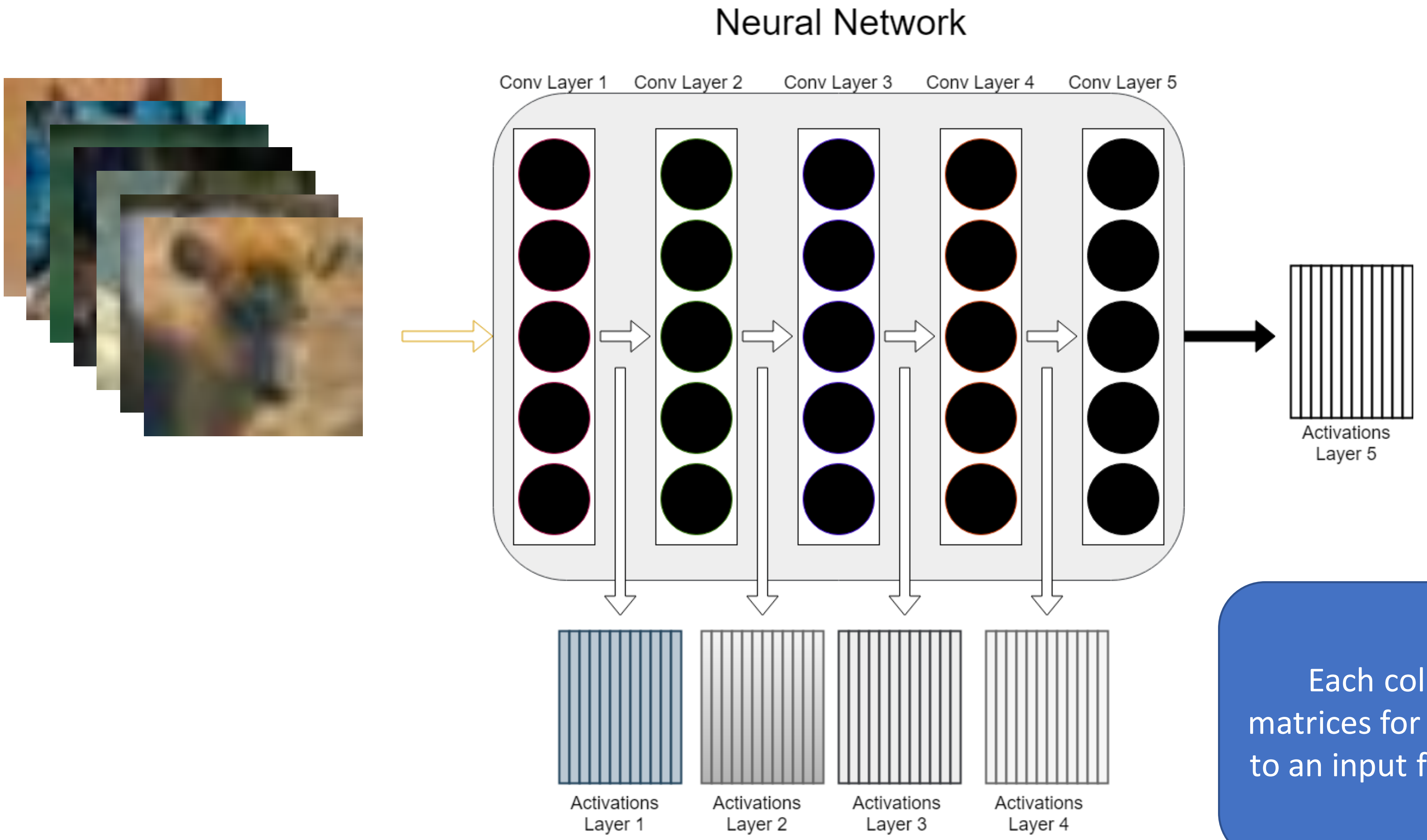
Background: Centered Kernel Alignment

$$CKA(X,Y) = \frac{HSIC(X,Y)}{\sqrt{HSIC(X,X)HSIC(Y,Y)}}$$

$$HSIC(X,Y) = \frac{trace(HXHHYH)}{(N-1)^2}$$

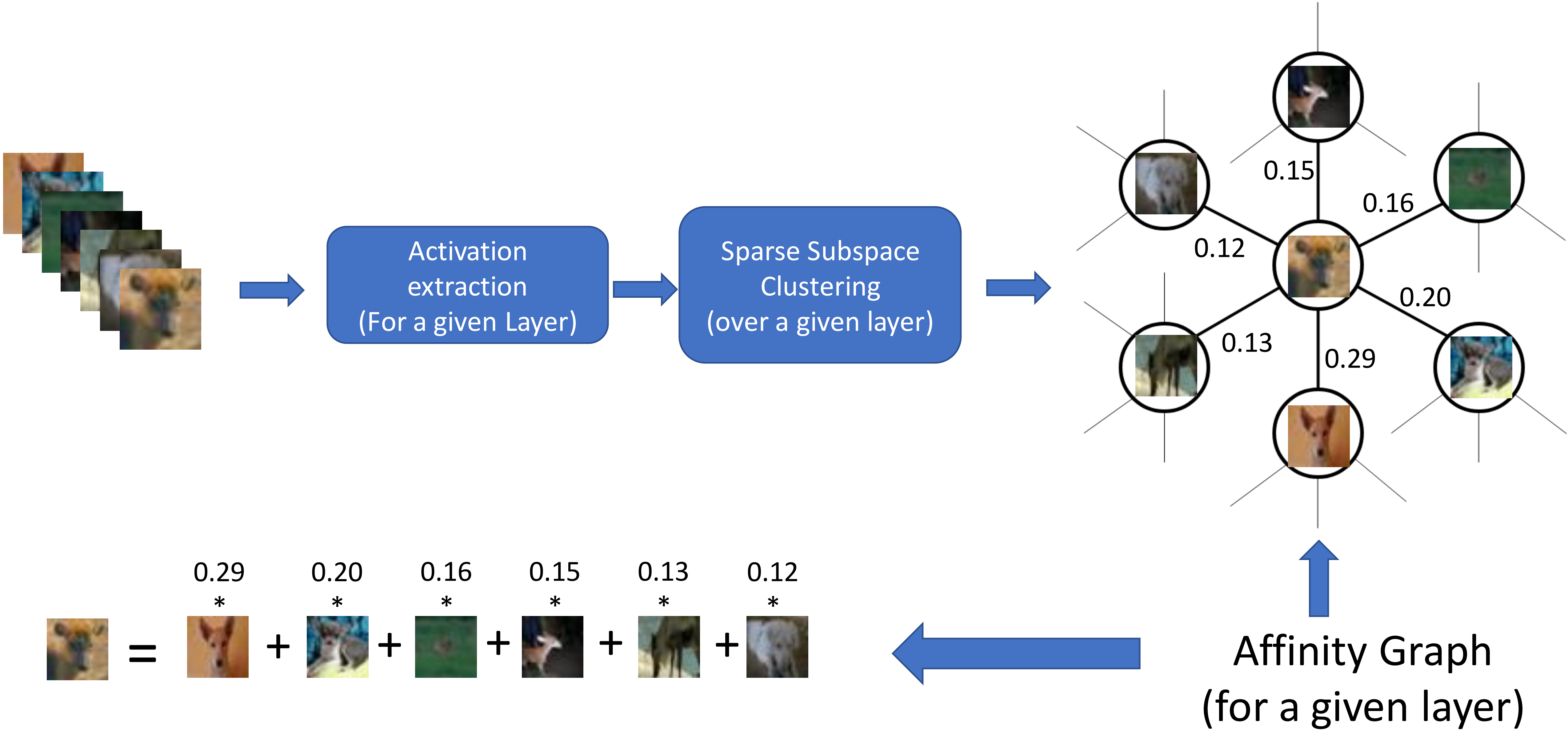
$$H = I - \tfrac{1}{N}\mathbf{1}\mathbf{1}^T$$

Network Analysis Setup: Activation Extraction



Each column of the Activation matrices for a given layer corresponds to an input fed to the neural network.

SSC Output


Activation extraction (For a given Layer)
Sparse Subspace Clustering (over a given layer)
0.15
0.16
0.12
0.20
0.13
0.29
0.29 *
0.20 *
0.16 *
0.15 *
0.13 *
0.12 *
=
+
+
+
+
+
Affinity Graph (for a given layer)

Experimental Analysis

- Having obtained the Layer-wise SSC Affinity Graphs, we branch out into 3 experimental trajectories.

Analyzing the community structure and convergence of the layer wise Affinity Graphs.

Analyzing Similarity between all layers wise Affinity Graphs using CKA.

Interpreting Neural Networks by analyzing their subspaces

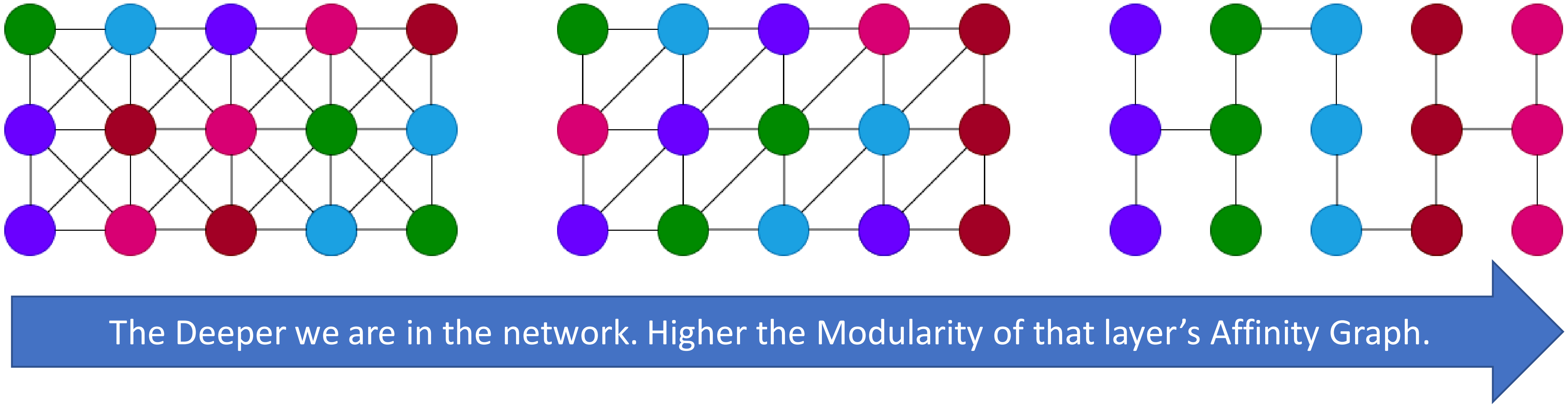
Layer wise SSC Affinity Graph Analysis

- We motivated SSC with an aim to learn affinity graphs.
- Now we focus on analyzing the community clusters in these layer-wise graphs.
- The coherence of the community structure in the graph is quantified by its Modularity Score :

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j)$$

Key Observation(i)

- <u>**The deeper we get into the network, the more homogenous the communities of that layer's affinity graph become.**</u>
- This Phenomena is quantified by the modularity score of the respective graphs.

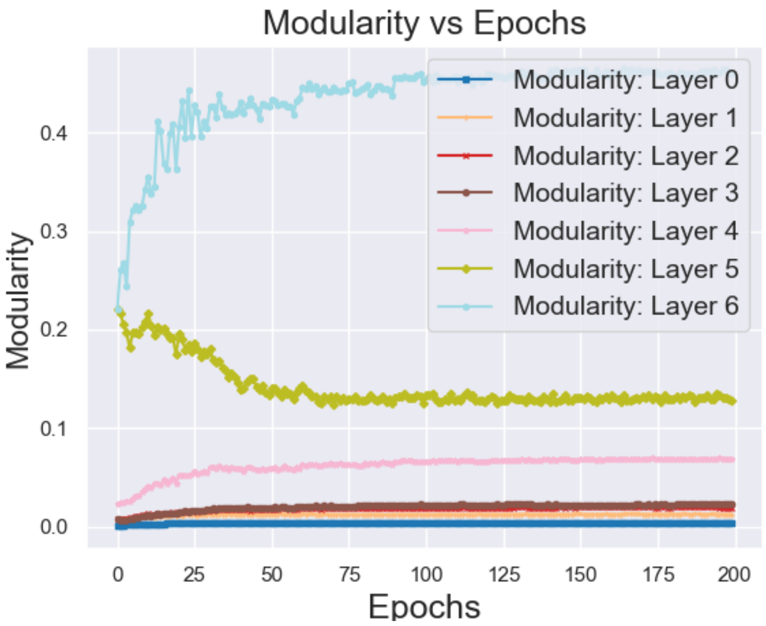


Layer wise SSC Affinity Graphs

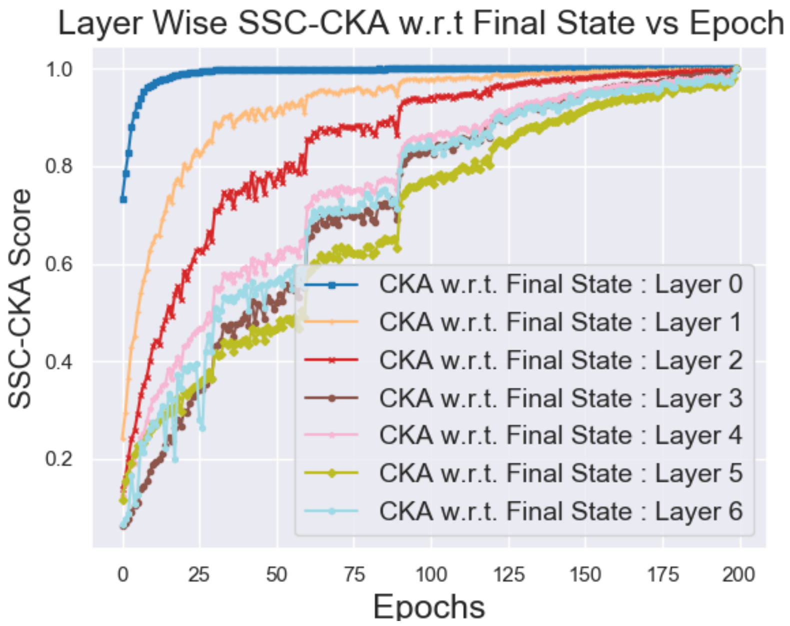
Key Observations(ii)

- We also observe that **deeper layers take more epochs to converge to their final representations.**
- This is quantified by CKA similarity of an Affinity Graph at a given epoch w.r.t. to its final state.
- The results are presented next.

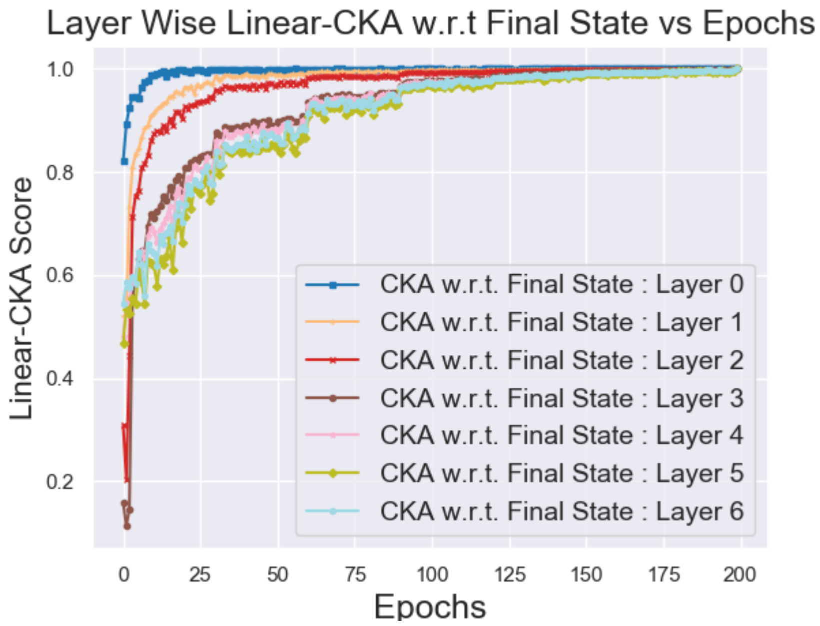
Analyzing Network Training Dynamics (1)



Modularity



SSC-CKA Convergence



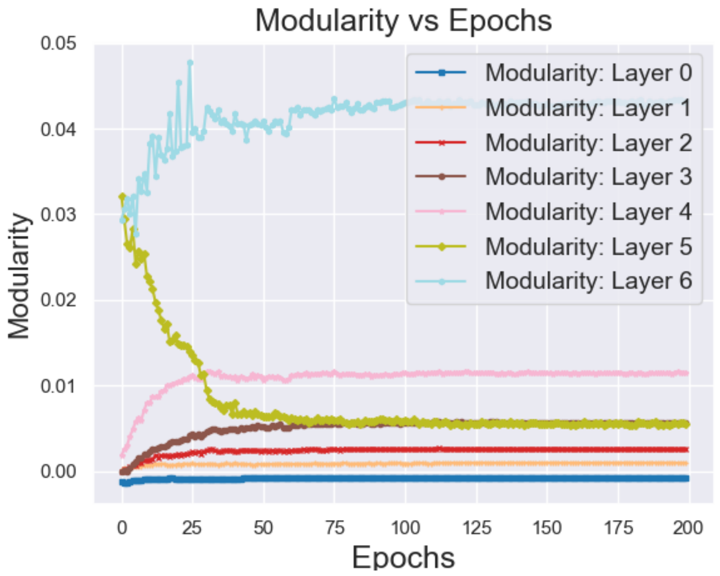
Linear-CKA Convergence

Network : Wide ResNet-28x10

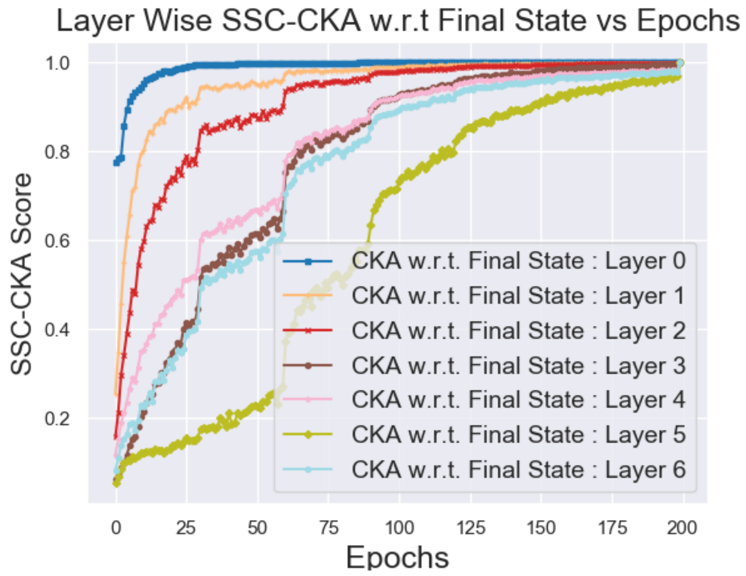
Dataset : CIFAR-10

Analyzing Network Training Dynamics (2)

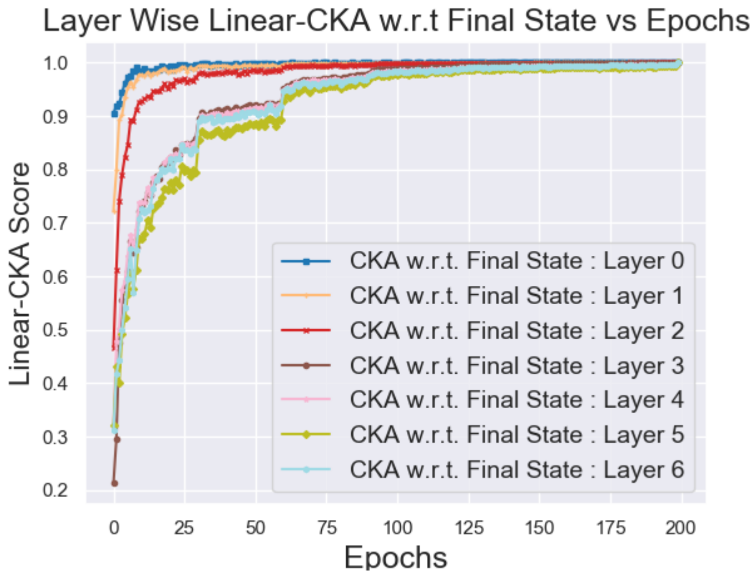
Modularity



SSC-CKA Convergence



Linear-CKA Convergence

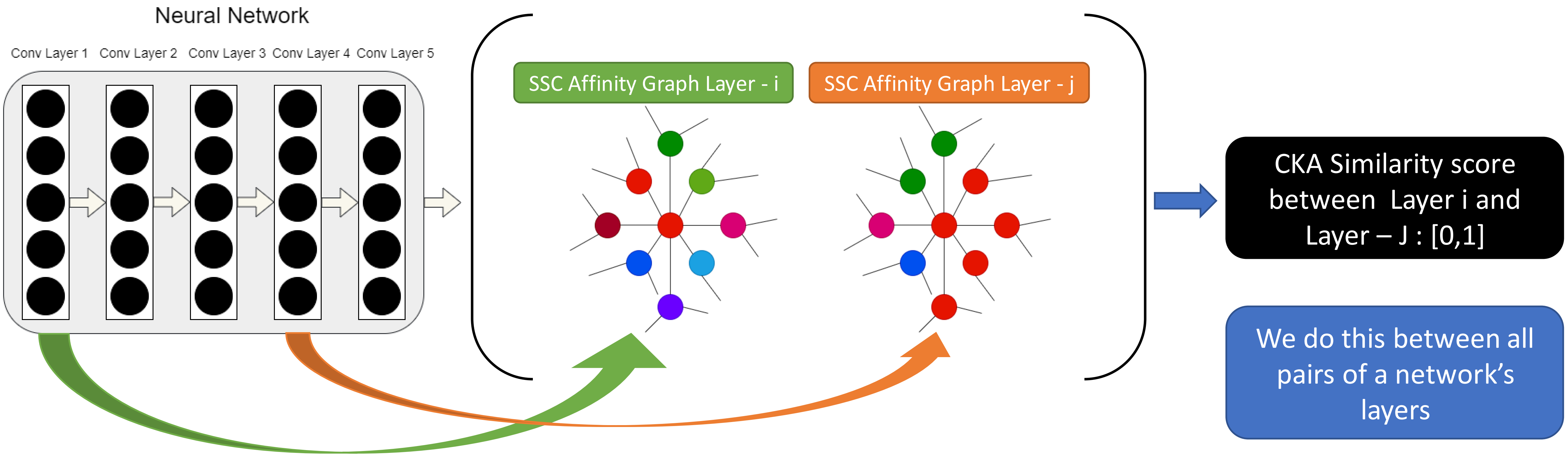


Network : Wide ResNet-28x10

Dataset : CIFAR-100

SSC-CKA Based Architecture Analysis

- We combine SSC with CKA to create tools that analyse Network Architectures.
- We do so by a pairwise comparison of all affinity graphs obtained by analysing all the convolutional layers of a neural network.

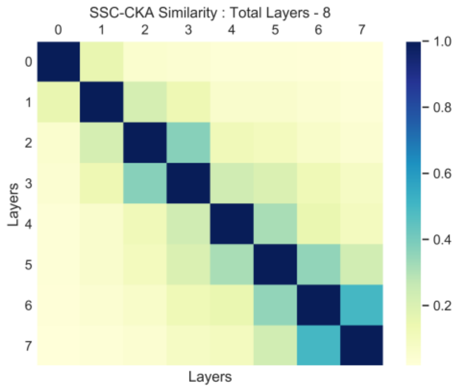


Aspects of Architectural Analysis

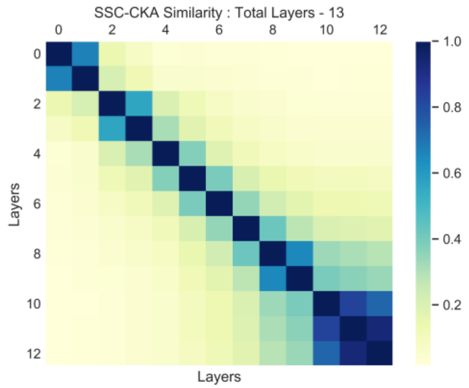
As a part of this analysis based on SSC-CKA, We Study the :-

1. Effects of Depth on the Layers of a Neural Network.
2. Effects of Width on the Layers of a Neural Network.
3. Effects of Epochs on the Layers of a Neural Network.
4. Effects of Data Quantity on the Layers of a Neural Network.

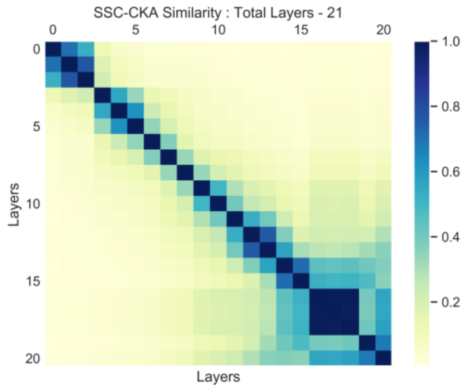
Effects of Depth – (1)



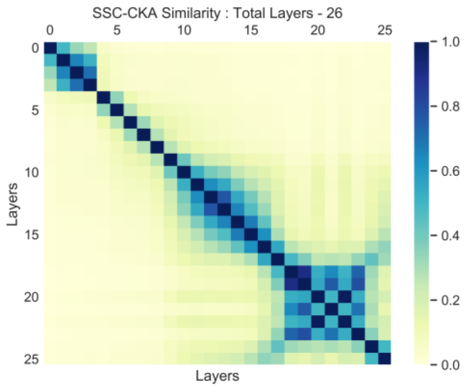
VGG-11 (92%)



VGG-16 (94%)



VGG-24 (93%)

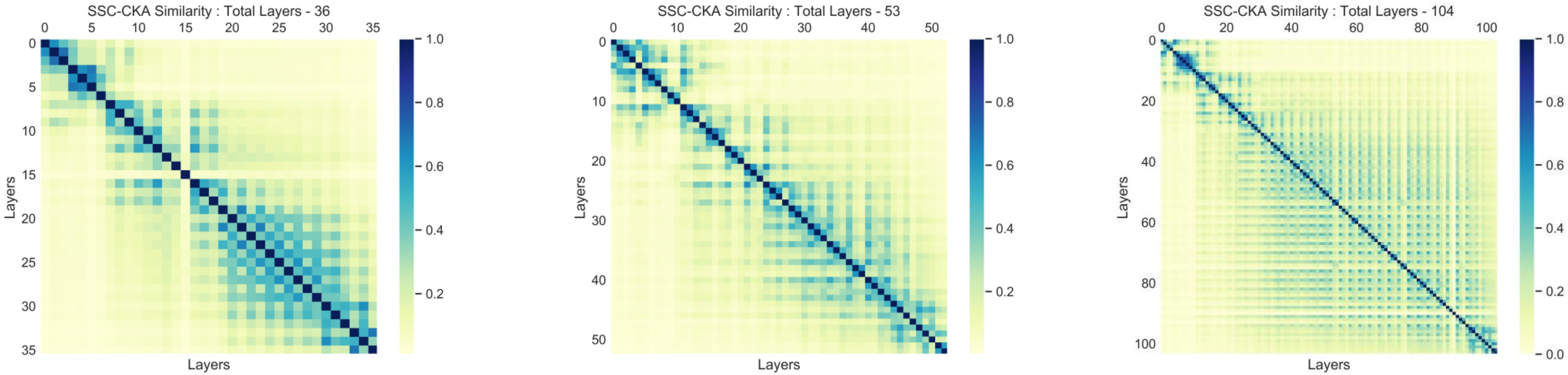


VGG-29 (62%)

As the depth of the network increases, so does block diagonal structure in the pairwise heatmap

Dataset : CIFAR-10

Effects of Depth – (2)



ResNet-36 (94%)

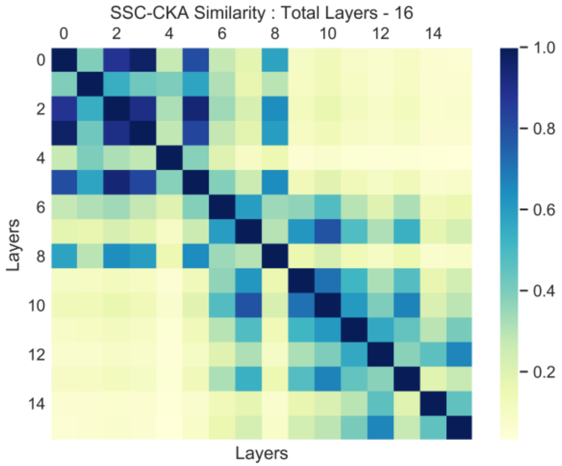
ResNet-53 (94%)

ResNet-104 (93%)

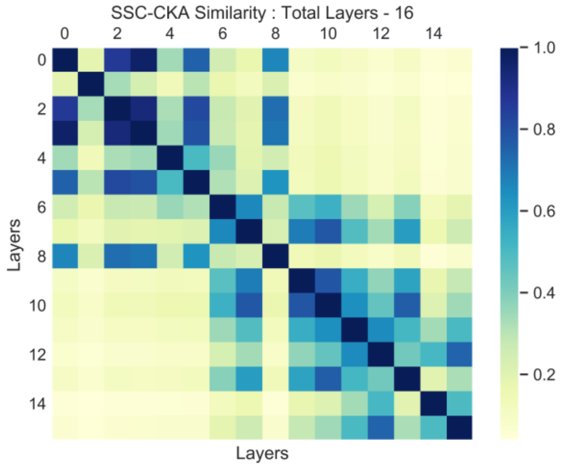
As the depth of the network increases, so does block diagonal structure in the pairwise heatmap

Dataset : CIFAR-10

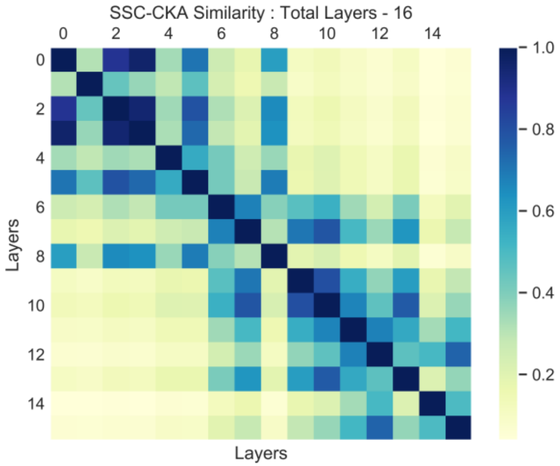
Effects of Width – (1)



Wide ResNet-16 – 2x (92%)



Wide ResNet-16 – 6x (93%)

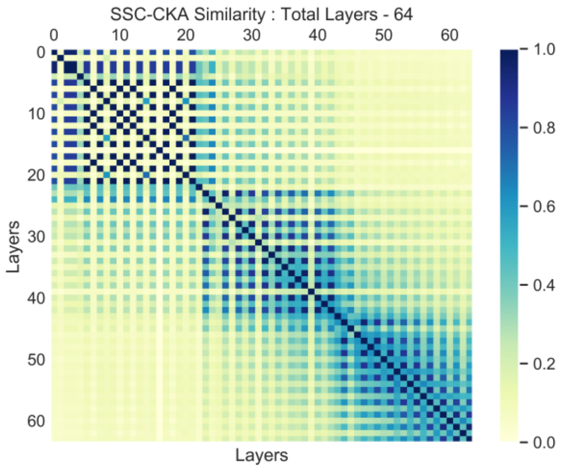


Wide ResNet-16 – 10x (94%)

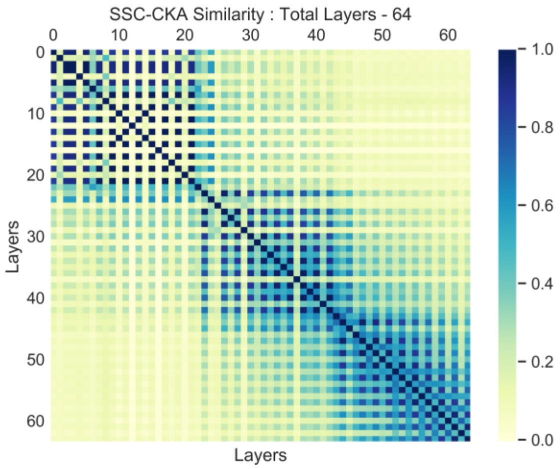
Block Diagonal Structure is Independent of the width of the Network

Dataset : CIFAR-10

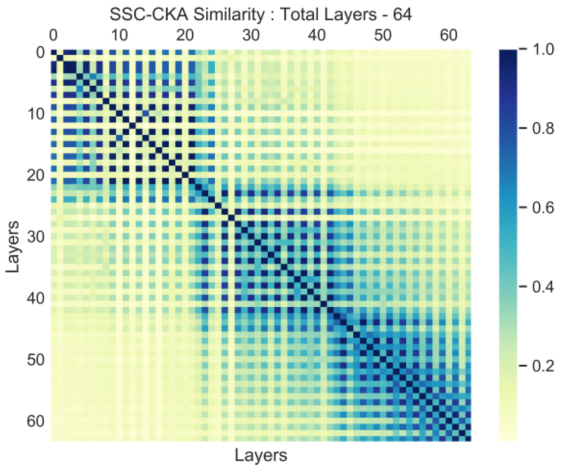
Effects of Width – (2)



Wide ResNet-64 – 2x (95%)



Wide ResNet-64 – 6x (95%)

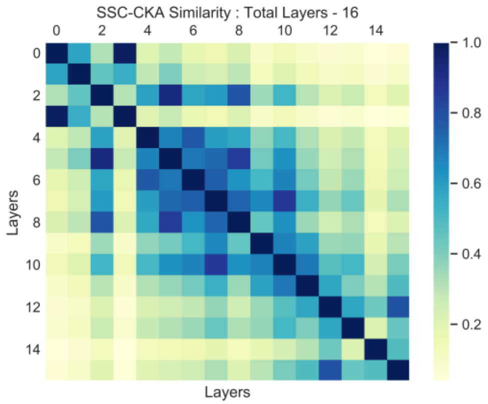


Wide ResNet-64 – 10x (96%)

Block Diagonal Structure is Independent of the width of the Network

Dataset : CIFAR-10

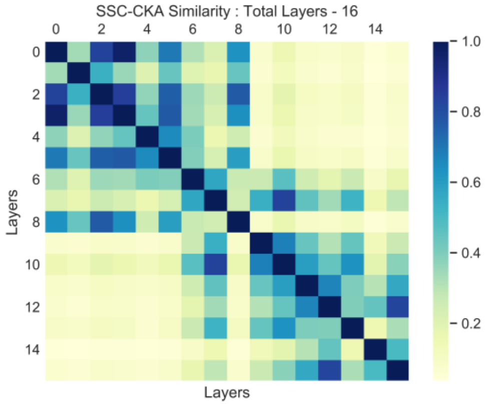
Effect of Epochs (1)



ResNet-16x2

Epoch 1

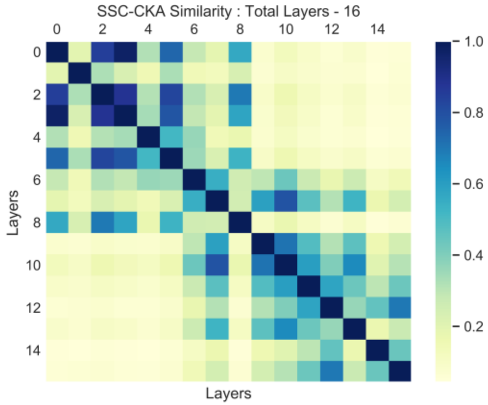
Accuracy 36%



ResNet-16x2

Epoch 30

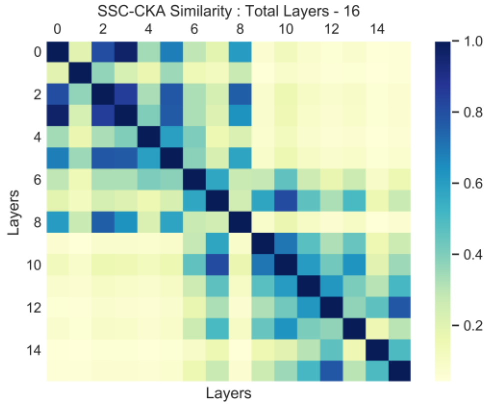
Accuracy 82%



ResNet-16x2

Epoch 60

Accuracy 91%



ResNet-16x2

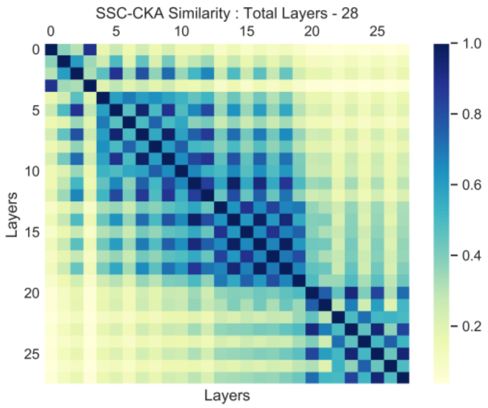
Epoch 100

Accuracy 92%

Block Diagonal Structure reduces as the training epochs Increase

Dataset : CIFAR-10

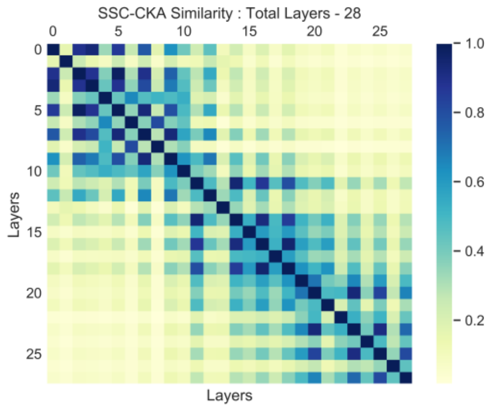
Effect of Epochs (2)



ResNet-28x2

Epoch 1

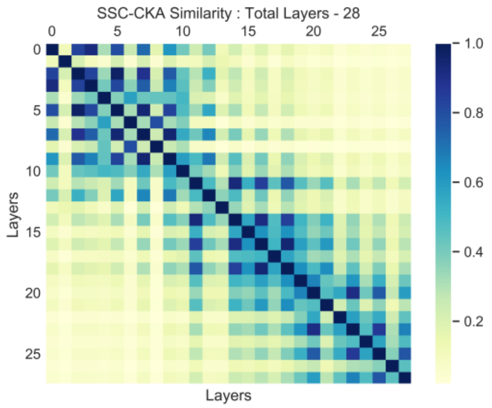
Accuracy 36%



ResNet-28x2

Epoch 30

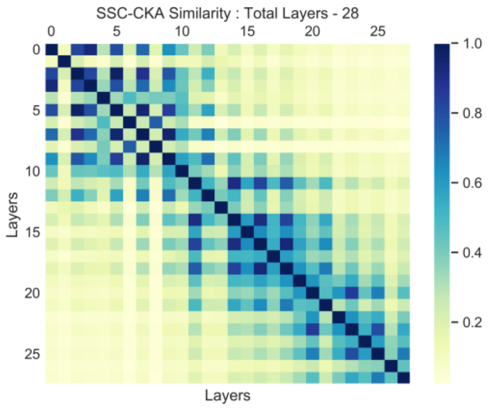
Accuracy 82%



ResNet-28x2

Epoch 60

Accuracy 91%



ResNet-28x2

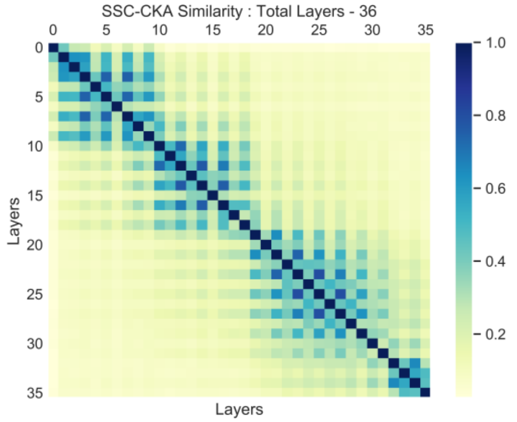
Epoch 100

Accuracy 92%

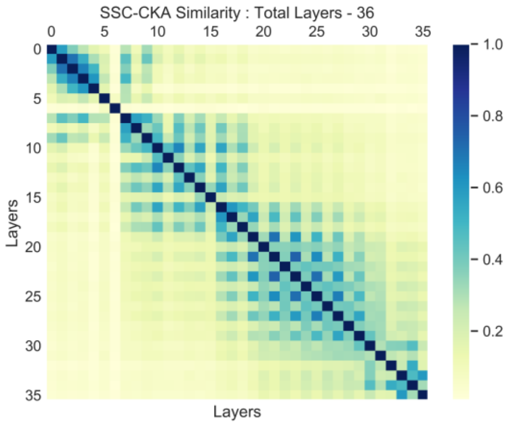
Block Diagonal Structure reduces as the training epochs Increase

Dataset : CIFAR-10

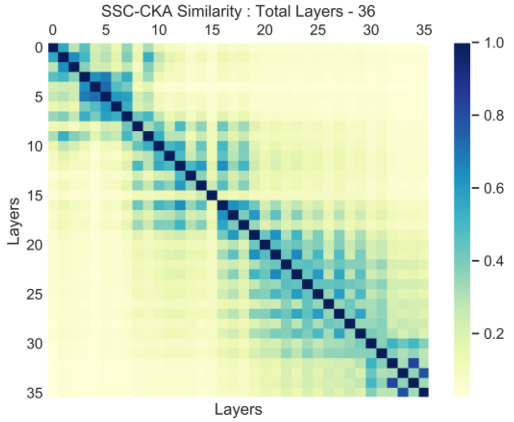
Effects of Training Data Quantity



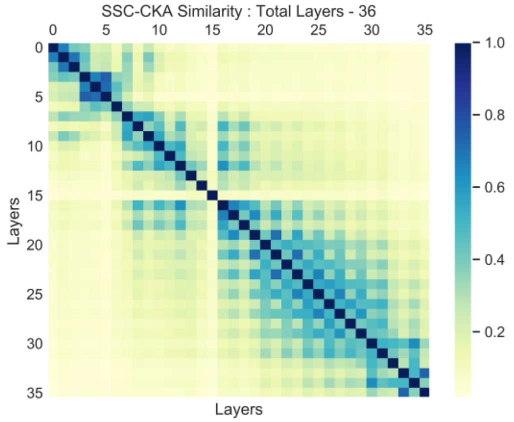
ResNet-34
5% training Data
56% Accuracy



ResNet-34
25% training Data
88% Accuracy



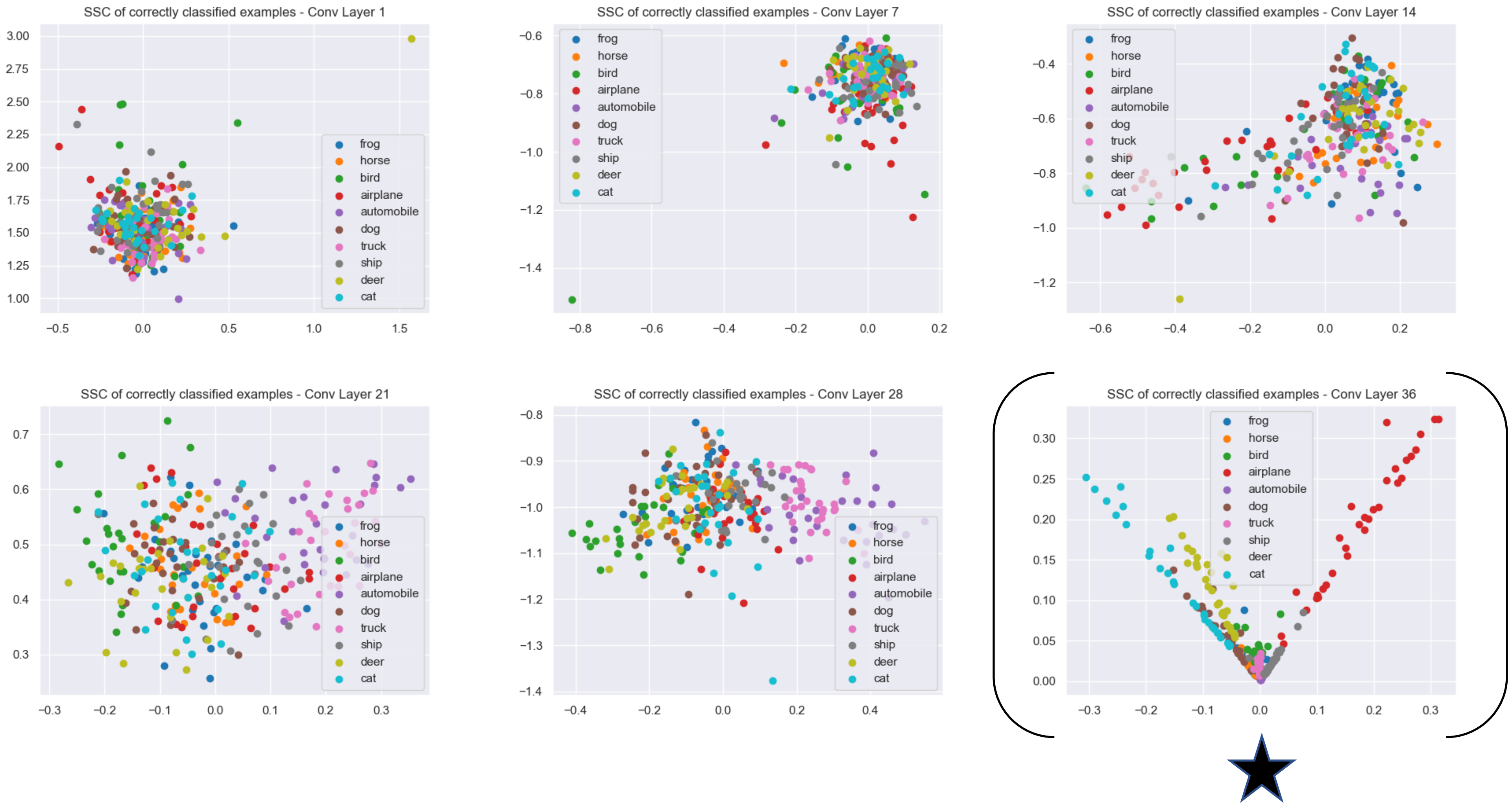
ResNet-34
50% training Data
92% Accuracy



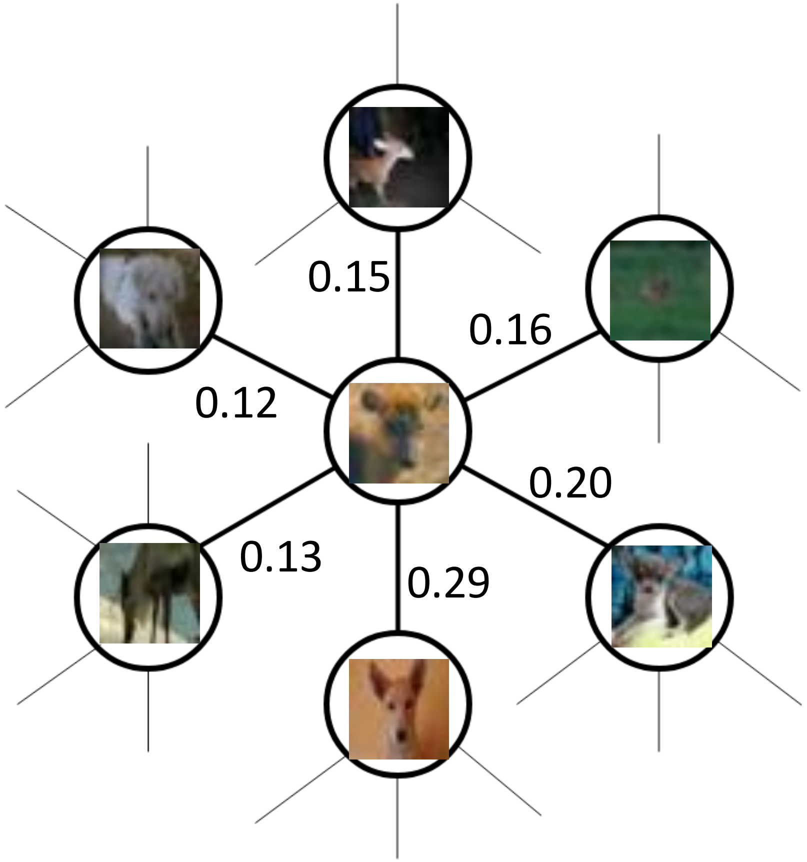
ResNet-34
100% training Data
94% Accuracy

Dataset : CIFAR-10

Latent Space Visualization (Correctly Classified Inputs)



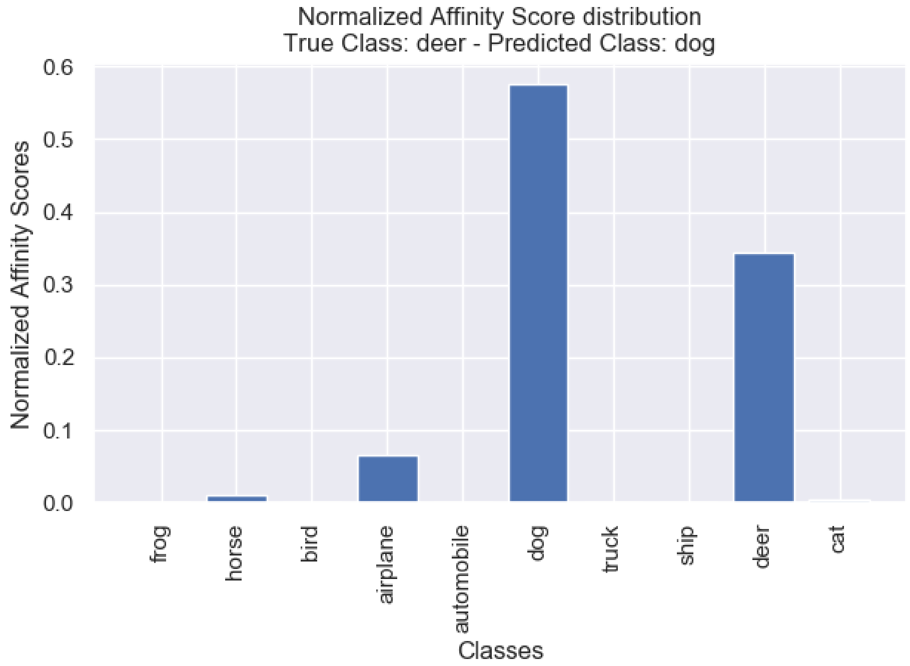
Interpreting Local Neighbourhood of Inputs

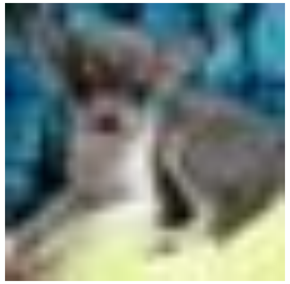
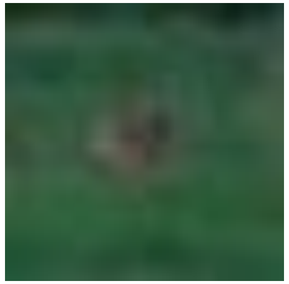


- Next, we analyse the local neighbourhoods in the learned Affinity graph for the final convolutional layer of the neural network.
- We do this for a few examples where the network made an **incorrect prediction with a high confidence**.
- We demonstrate the ability of SSC to quantify the local Neighbourhood of an input and demonstrate its correlation with the network's prediction.
- We also demonstrate certain pairwise comparisons that help contextualize the network's prediction.

Instance Neighbourhood Visualization (i)

Original : Deer
Classified : Dog
Classification confidence : 78%




Highest Affinity Neighbours in the Neighbourhood Affinity graph
Affinity Descending from Left to Right

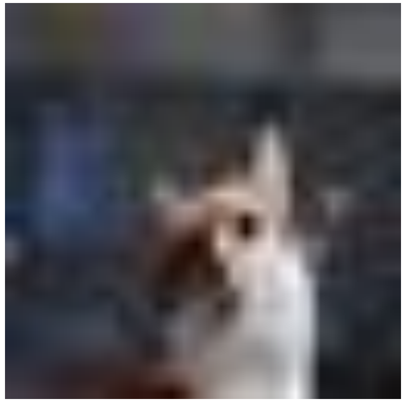
SSC-Label and Network Label Correlation

Network	SSC Prediction	Network Prediction	Correlation
ResNet-36	95.8%	95.8%	**98.3%**
ResNet-18	95.4%	96.1%	**97%**
ResNet-53	90.6%	94.8%	**93.5%**

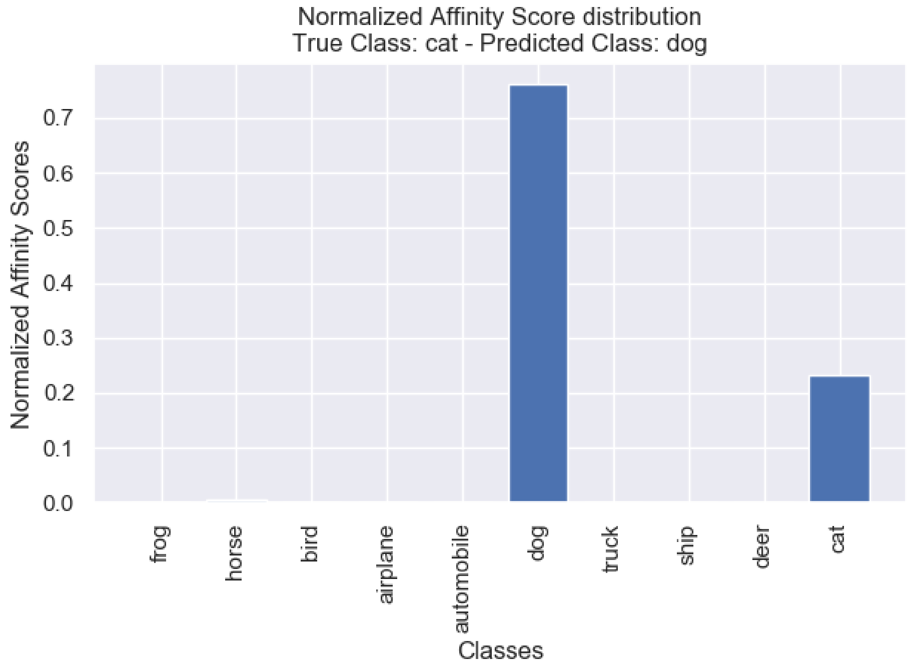
SSC Prediction refers to the highest weighted class in the Neighbourhood of a given input. (Layer : Final Convolution Layer)

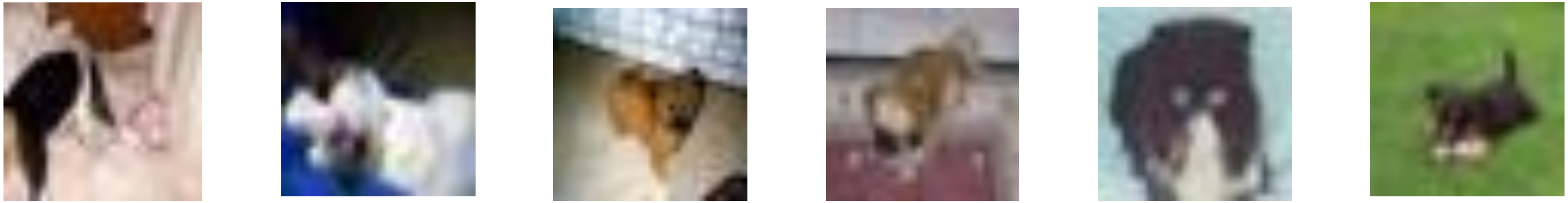
High correlation between network output and SSC at final convolution layer suggests that network is trying to separate data into a disjoint union of subspaces.

Instance Neighbourhood Visualization (ii)



Original : Cat
Classified : Dog
Classification confidence : 94%

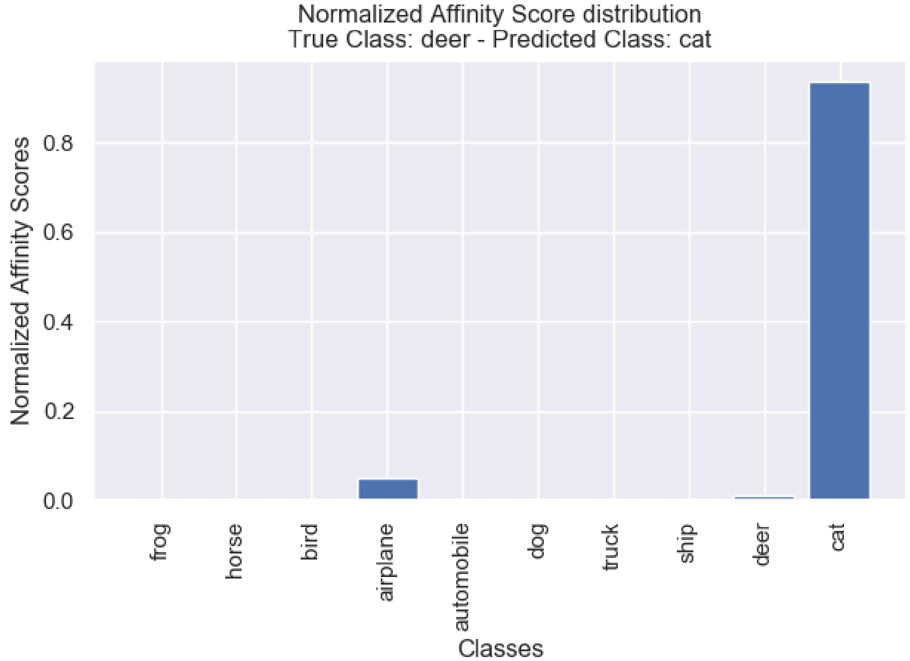




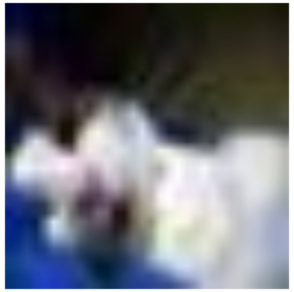
Highest Affinity Neighbours in the Neighbourhood Affinity graph
Affinity Descending from Left to Right

Instance Neighbourhood Visualization (iii)

Original : Deer
Classified : Cat
Classification confidence : 98%



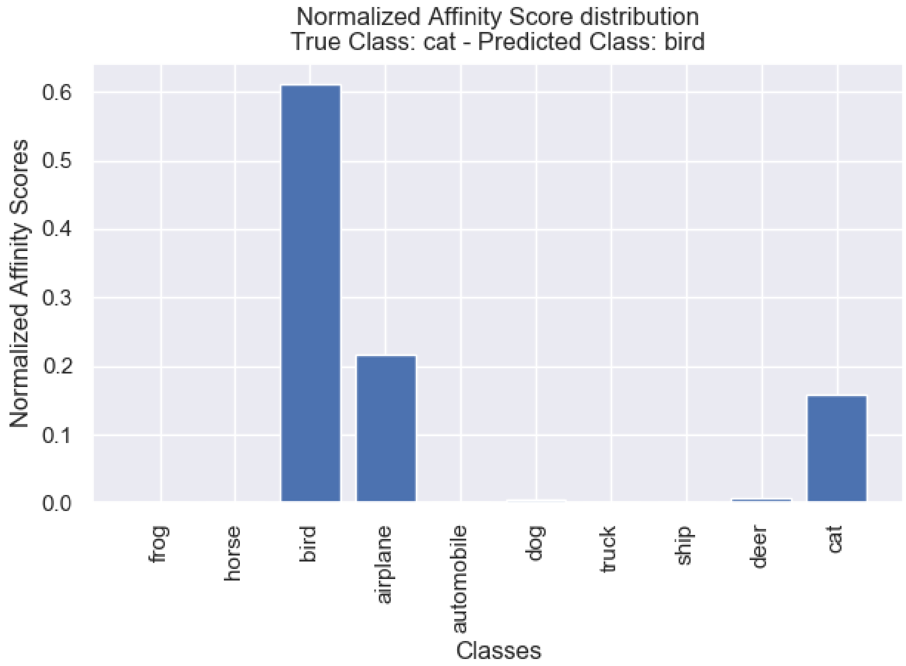


Highest Affinity Neighbours in the Neighbourhood Affinity graph
Affinity Descending from Left to Right

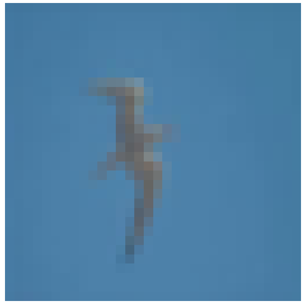
Instance Neighbourhood Visualization (iv)

Original : Cat
Classified : Bird
Classification confidence : 97%







Highest Affinity Neighbours in the Neighbourhood Affinity graph
Affinity Descending from Left to Right

Thank you for your Time and Attention

Code : https://github.com/23Uday/Subspace-Clustering-based-analysis-of-Neural-Networks