Literature Review - 1

- Narasimha Prasanth Chintarlapalli Reddy
  - Narasimha_Reddy@student.uml.edu
  - 01669930

The **primary** research paper I chose for this Literature Review is **"Visualizing the Hidden Activity of Artificial Neural Networks "**[1] by **Paulo E. Rauber, Samuel G. Fadel, Alexandre X. Falcao, and Alexandru C. Telea**.

@ARTICLE{7539329,
author={P. E. Rauber and S. G. Fadel and A. X. Falcão and A. C. Telea},
journal={IEEE Transactions on Visualization and Computer Graphics},
title={Visualizing the Hidden Activity of Artificial Neural Networks},
year={2017},
volume={23},
number={1},
pages={101-110},
keywords={data visualisation;learning (artificial intelligence);neural nets;pattern classification;artificial neural networks;dimensionality reduction;hidden activity visualization;high-dimensional vectors;machine learning;pattern classification;Benchmark testing;Computational modeling;Data visualization;Neural networks;Neurons;Training;Visualization;Artificial neural networks;algorithm understanding;dimensionality reduction},
doi={10.1109/TVCG.2016.2598838},
ISSN={1077-2626},
month={Jan},}

The **secondary** research paper I chose is **"Understanding deep features with computer-generated imagery "**[2] by **M. Aubry and B. Russell**.

@INPROCEEDINGS{7410686,
author={M. Aubry and B. C. Russell},
booktitle={2015 IEEE International Conference on Computer Vision (ICCV)},
title={Understanding Deep Features with Computer-Generated Imagery},
year={2015},
volume={},
number={},
pages={2875-2883},
keywords={feature extraction;image colour analysis;image representation;neural nets;principal component analysis;3D CAD models;3D viewpoint;CNN feature analysis;computer generated imagery;convolutional neural networks;image datasets;input scene factors;linear decomposition;network representation;object categories;object style;principal component analysis;rendered images;scene lighting configuration;understanding deep features;Computational modeling;Feature extraction;Lighting;Principal component analysis;Rendering (computer graphics);Solid modeling;Three-dimensional displays},

doi={10.1109/ICCV.2015.329},
ISSN={},
month={Dec},}

Let's start by analyzing the secondary research paper. The paper introduces an approach for analyzing the features generated by convolutional neural networks(CNNs). It includes many factors(features) such as object style, 3D viewpoint, color, and scene lighting. The datasets used for this paper are:

1. AlexNet[3]
2. Places[4]
3. Oxford VGG[5]

This paper demonstrates that the analysis based on computer-generated imagery translates to the network representation of natural images. The approach to analyze the images can be summarized as follows. The minimal input for this analysis is a set of related images. First, their features are analyzed jointly and this only gives high level information as it cannot identify the origin of the variation of the input images. The factors influencing this can be any of the above mentioned features. Now, it focuses on computer generated images and have full control over different factors. By representing, analyzing them separately and comparing their relative importance, it can learn the influence of different factors individually. The paper calls the analyzing features jointly as 'Image Collection Analysis' and individual analysis of factors as 'Multiple Factor Analysis'.

This paper uses Principal Component Analysis(PCA) to analyze the factors which makes sense because the PCA algorithms puts emphasis on most important features. The following images summarizes the results obtained from the approach used.

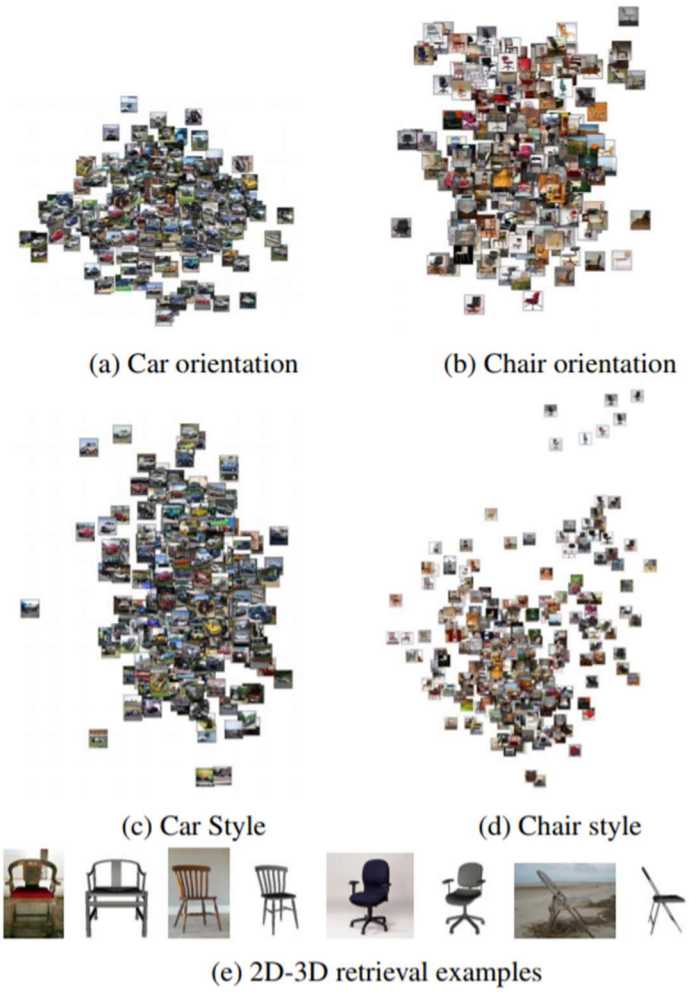

Figure:PCA embeddings over AlexNet pool5 features for cars and chairs with orientation and style separated.

In the above figure, we can see how style and orientation factors differ for car and chair. The paper also produces results for influence of each of ViewPoint, Style, and lighting for all the three datasets mentioned. Thus, the approach can be used to identify the importance of each factor individually.

Coming to our primary paper, it's focus is mainly on learning the importance and influence of hidden layers in a Neural Network. This paper has two main objectives. They are exploring the relation between alternative representations of observed Artificial Neural Networks(Let's call this T1) and exploring the relation between artificial neurons or hidden layers(Let's call this objective T2). The datasets used for this paper are:

1. MNIST[6]
2. SVHN[7]
3. CIFAR-10[8]

Two kinds of ANNs are considered in this paper. Multilayer perceptrons (MLPs) and convolutional neural networks (CNNs). While there are larger models than these two, these are sufficient for the datasets mentioned.

Learning the relation between alternative representations of observed ANNs(T1) can be achieved by learning the relations in each and every layer and comparing them with the next layer. The paper first understands what untrained ANNs know and then after training the data, it compares the classes in each and every layer. For MNIST database, the paper compared untrained ANN with last MLP hidden layer after 1000 activations.

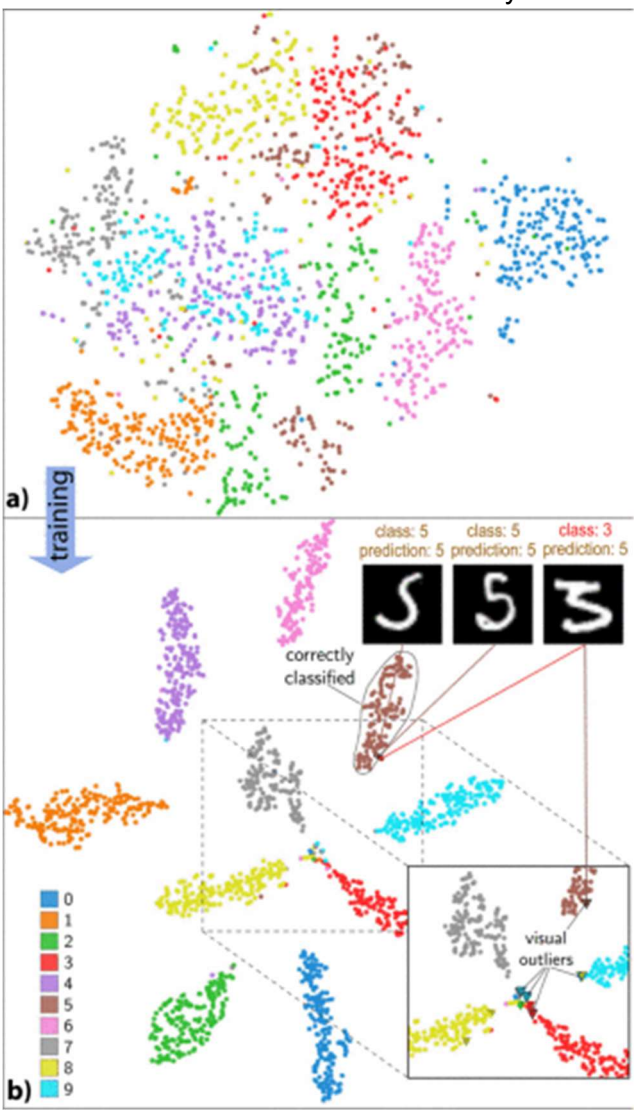

Figure: Projection of the last MLP hidden layer activations, MNIST test subset. a) Before training . b) After training. Inset shows classification of visual outliers.

The figure above which is taken from the paper is a good visual representation to understand how trained clusters differ from untrained ANNs. The similar process is carried out with the other two datasets. It is a little complex with the other two datasets as explained in the paper.

To achieve our second objective T2 i.e, the relationship between hidden layers, we need to analyze and understand each and every hidden layer. The above process tells us how the clusters are classified before and after training. However, It does not tell us the relation between the hidden layers. So, the paper defines dissimilarity $d_{i,j}$ between neurons i and j as $d_{i,j}=1-|r_{i,j}|$, where $r_{i,j}$ is the empirical (Pearson's) correlation coefficient between neurons i and j on a dataset composed of layer-1 activations. This metric is very useful as we can define both negative and positive relations between layers.

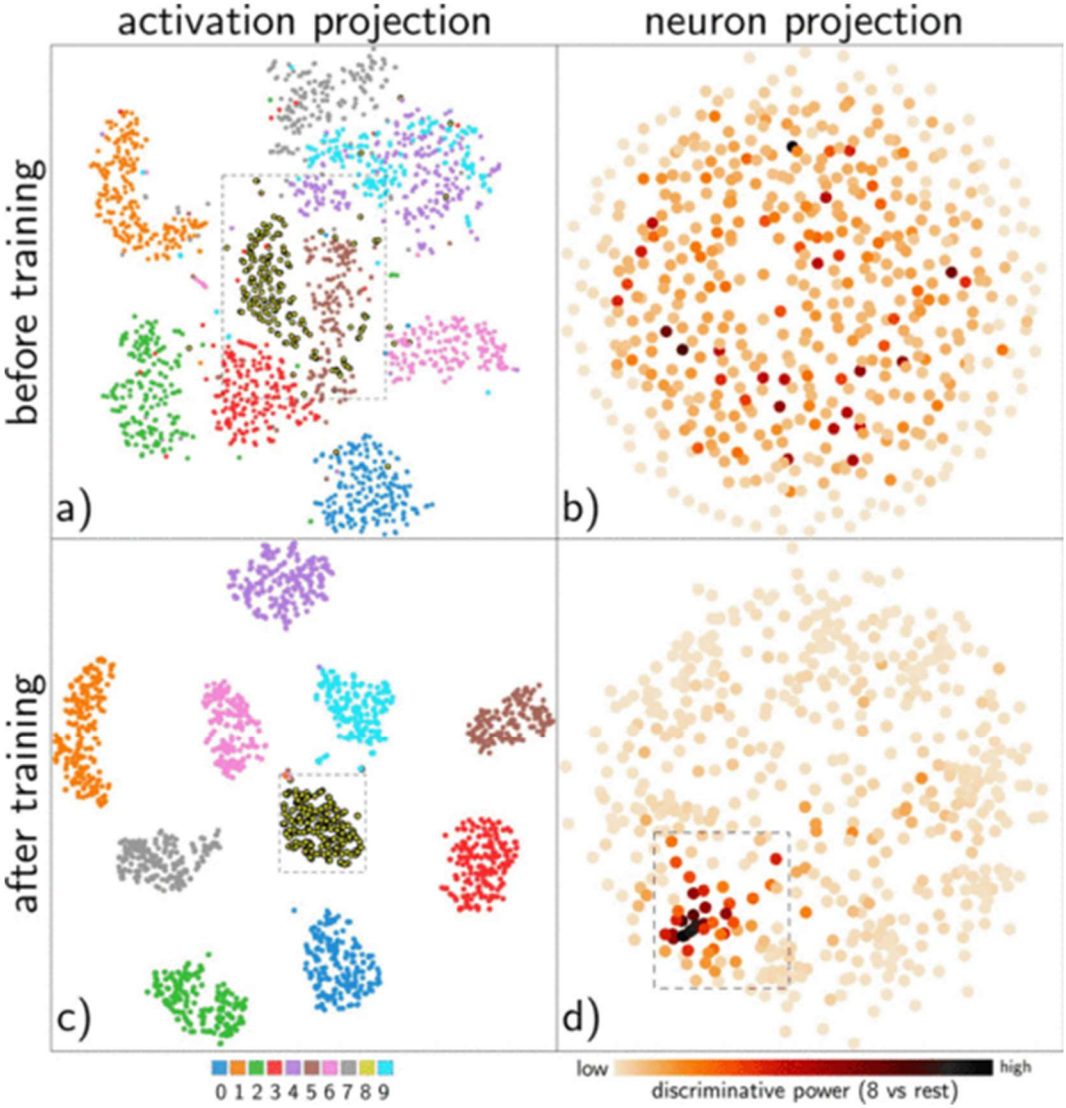

Figure: Activation and neuron projections of last CNN hidden layer activations before and after training, *MNIST* test subset. Neuron projection colors show the neurons' power to discriminate class 8 *vs* rest.

In the above figure, we can clearly see(b) that before training, the neurons with good discriminative power are scattered and with proper training, all the neurons with good discriminative power work together collectively for better results.

The above observation should be considered very important since we can learn the relation between hidden layers which can be very useful.

Comparing both papers:

The secondary paper deals with analyzing how each factor influences the final outcome individually. This is a very important concept in decision making. It is always helpful to know which feature has most importance. The primary paper, however, deals with learning the importance of each layer. The secondary paper's observations are very useful in achieving primary paper's T1 objective. Only if we knew importance of each factor, we can learn the relationship between trained and untrained ANN. **The secondary paper's observations can be used to improve decision making while the primary paper's observations can be used in writing a better algorithm for CNN and MLP.** The primary paper claims that it is the first paper to analyze a Neural Network as explained, which opens for a whole new world of possibilities of utilizing and improving the concept.

References:

[1] P. E. Rauber, S. G. Fadel, A. X. Falcão and A. C. Telea, "Visualizing the Hidden Activity of Artificial Neural Networks," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 101-110, Jan. 2017.
doi: 10.1109/TVCG.2016.2598838

[2] Mathieu Aubry and (2015). Understanding deep features with computer-generated imagery. *CoRR, abs/1506.01151*, .

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In NIPS, 2012.

[4] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using Places database. In NIPS, 2014.

[5] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In Proc. BMVC., 2014.

[6] Y. LeCun, C. Cortes, C. J. Burges, The MNIST database of handwritten digits, 1998, [online] Available: http://vann.lecun.com/exdb/mnist/.

[7] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A. Y. Ng, "Reading digits in natural images with unsupervised feature learning", Proc. Neural Information Processing Systems, vol. 2011, pp. 5, 2011.

[8] A. Krizhevsky, G. Hinton, Learning multiple layers of features from tiny images, 2009, [online] Available: www.cs.toronto.edu/-krizz/learning-features-2009-TR.pdf