Unsupervised clustering based understanding of CNN

Deeptha Girish girishde@mail.uc.edu
Vineeta Singh singhvi@mail.uc.edu
Anca Ralescu anca.ralescu@uc.edu
University of Cincinnati

Abstract

Convolutional Neural networks have been very successful for most computer vision tasks such as image recognition, classification, object detection and segmentation. Even though CNNs are very successful and give superior results as compared to traditional image processing algorithms, interpretability of their results remains an important issue to be solved. Indeed, lack of interpretability and explainability of how CNN work at their various levels, caused a certain skepticism among their potential users, as for example those working in medical diagnosis or autonomous driving cars. The current study aims to answer some of the issues related to interpretability by the use unsupervised methods to discern the features learned by the CNN in different layers.

1. Introduction and Background

CNN visualization techniques have allowed us to visualize the weights learned at different layers. It has been observed that the initial layers learn basic low level image features such as edges while the upper layers learn more complicated features such as shapes etc.

Many visualization techniques have been proposed in the past to understand the working of convolutional neural networks. Just visualizing CNN activations have shown us that the lower layers of the CNN learn low-level image features such as edges and colors. CNN activations of higher level layers provide no insight on what exactly is being learnt. A more sophisticated visualization by Zeiler et al. [9], is the use of occluders on the image. This, in conjunction with a classification procedure helps understand just which parts of the image lead to classification of the image into the right class.

CNN codes (the activations of the layer in a CNN before classification, including non-linearity) capture a lot of information about the image and have worked well as features for images used in many classification tasks. This work takes a step further in investigating the response of the individual layers to images of different classes.

2. Algorithm

Initially, in order to understand what kind of features are learnt in every layer, we devised the following experimental procedure.

1. Select n the number of clusters/classes.
2. Select, from the ImageNet dataset, subsets of equal size k, from each class. Thus in total there are nk images.
3. Each image is passed through the pre-trained network and their activations for all layers is recorded: thus at layer i there are nk activations, which constitute the dataset D_i, for the analysis at that layer described in the next step.
4. At layer i, the dataset D_i is clustered into *n* clusters using the k-means algorithm which is explained below.
5. Analyze the clusters obtained at each layer with respect to the original classes to which the images they correspond to belong.

The K-means algorithm [2] is used for clustering the layer activations. The K-means algorithm stores k centroids that it uses to define clusters. A point is considered to be in a particular cluster if it is closer to that cluster's centroid than any other centroid.

K-Means finds the best centroids by alternating between (1) assigning data points to clusters based on the current centroids (2) chosing centroids (points which are the center of a cluster) based on the current assignment of data points to clusters [5].

3. Experimental Results

We applied the algorithm described above as follows.

The first experiment was carried out with the Alexnet [4] network pre-trained on the imagenet dataset [6], following the algorithm steps indicated above. It was seen through our experiments that for the initial layers, there is no class specific pattern observed for any cluster: the clusters had a random mix of images from all classes. However, the clusters

formed from higher layers better captured classes and therefore the features captured were more class specific: images from the same class were clustered together. Figure 1 shows the layers present in AlexNet. The clusters formed from

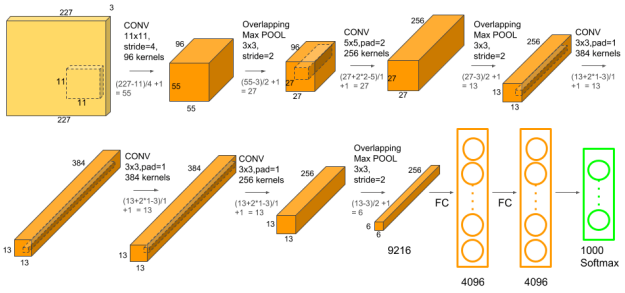

Figure 1. AlexNet [1]

the conv5 layer activations (the last convolutional layer of Alexnet) had minimum number of mis-classifications. Almost all images of the same class were clustered together.

This suggests that the earlier layers learn basic low-level features that are common to images of all classes. Due to this, the clusters formed from the activations of the first few layers are random. In other words, the clusters are formed based on basic image features common to all classes. Hence, no cluster belongs to a single class. By contrast, the clusters formed from higher layer activations revealed class identities: each cluster had majority images from one particular class, suggesting that the higher level layers learn more complicated class specific features. These features are, in a sense, implicit. They cannot be extracted directly from the image/pixel information. Only images from a particular class get activated because only they may have that feature, so when clustering is performed images of the same class get clustered together.

For initial experiments we used five classes: cars, dogs, flowers, chairs and mugs. Figure 2 depicts the result of clustering for activations extracted from each of the layers. We show the images present in a particular cluster for each one of the layers of the CNN (convolutional layer 1, convolutional layer 2, convolutional layer 3, convolutional layer 4, convolutional layer 5 , fully connected layer 6 and fully connected layer 7).

It can be seen from the table that the clusters in the lower layers (CONV1, CONV2 and CONV3) do not reveal class identities, which means that each of these clusters is a mixture of images from different classes. From this we can infer that the initial convolutional layers encode features that are common to images from all classes which could are basic low-level features such as edges, colors etc... We can see that in higher levels of the CNN (Fully connected layers; FC6 and FC7) the clusters reveal class indentities. Almost all images in the clusters belong to the same class. This means that the higher level features encode more complex features which might be specific to each class.

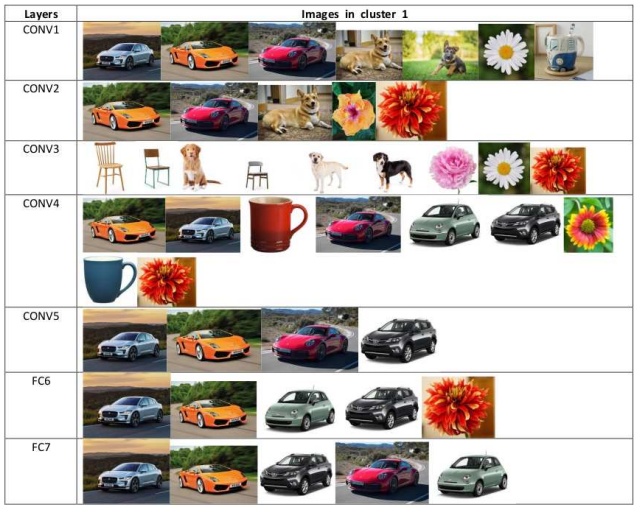

Figure 2. Images clustered based on CNN layer activations

A second experiment was run on deeper networks VGG16 and VGG19 [7] and observed the same trend.

4. Current work to be included in the full paper

In order to validate the approach described above, we are presently experimenting with the same algorithm with more complex network, such as ResNet [3] and InceptionNet [8]. Further more, we will use a larger dataset where class hierarchies are present. Using hierarchical clustering it might be possible to understand, to what extent the class hierarchies are captured across the CNN layers.

References

[1] Alexnet - imagenet classification with convolutional neural networks, Nov 2018.

[2] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[5] C. Piech. K means.

[6] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[7] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[9] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.