ISSN 2616-5775 Vol. 6, Issue 12: 118-124, DOI: 10.25236/AJCIS.2023.061213

Decoding the Black Box through a Comparative Study on Clustering Features in Convolutional Neural Networks

Ruiyi Zhang

Troy High School, Fullerton, California, America

Abstract: *Over the years, as Convolutional Neural Networks (CNNs) have revolutionized the field of image processing by achieving state-of-the-art results, questions on their internal working mechanics have arisen. As of this moment in time, the internal workings of CNNs remain a "black box", making it challenging to grasp and understand their decision-making processes. This study aims to unveil the mysteries of the black box by exploring how images are represented within CNNs. Accomplished with the TinyImageNet Data and VGG16 architecture, we extracted features from the penultimate layer and utilized K-means clustering to group the features. Through these clusters, we were able to uncover meaningful patterns and similarities among a vast amount of images, gaining insight into the internal mechanics of CNNs and the features they prioritize. Although parameters were adjusted to accommodate the purpose of understanding CNNs, a comparative analysis with other clustering methods is conducted to reveal more information about their efficacy and mechanics. This study has not only revealed more information about the internal workings of CNNs, but it also hopes to open the gates to more interpretable deep learning models in the future.*

Keywords: *Black Box, Convolutional Neural Networks*

1. Introduction

In the realm of image processing, the emergence of CNNs has been revolutionary, enabling a myriad of practical applications that were previously unapproachable. CNNs has demonstrated its potential by achieving state-of-the-art results on tasks of extreme difficulty from interpreting medical imaging to applications of autonomous driving. This type of deep learning model has been shown to have a prodigious ability in identifying and learning from patterns in vast amounts of data. Given said, CNNs have solidified their position as one of the leading areas in artificial intelligence research.

However, difficulties in regards to understanding their decision-making process parallel their astounding prowess. The intricate architecture of CNNs creates a "black box" that obscures its internal workings from human comprehension. This lack of transparency, which in terms creates a lack of understanding, is immensely detrimental to not only the applications that demand the understanding of why certain decisions were made but also the potential progress that could be made to CNNs themselves.

With the increasing demands of uncovering the black box, the academic landscape is replete with efforts aimed at interpreting and visualizing CNNs. Some existing research has involved techniques such as saliency maps [7], activation maximization [8], and feature inversion [9] to gain insights into these models. It is apparent that researchers are currently deeply engaged in understanding the nuances of image representation within CNNs. Hence, a comprehensive understanding remains uncharted. The current situation calls for a new methodology that places a heavier emphasis on a more high-level view of the image representations within CNNs.

This research intends to delve deeper into this enigma. We explore the internal representations of images within CNNs by leveraging the capabilities of K-Means clustering. The foundation of the study is based on the TinyImageNet dataset that is processed through the VGG16 model [1]- a choice encouraged by the dataset's diversity and the model's popularity within the research community.

The results yielded from this study intend to serve not only an academic curiosity but also pave the way for future research and progress in unraveling the internal working mechanics of CNNs with the purpose that deep learning models are powerful and transparent. Such advances will enable applications of CNNs to be more easily trained, debugged, and deployed ethically in real-world cases.

ISSN 2616-5775 Vol. 6, Issue 12: 118-124, DOI: 10.25236/AJCIS.2023.061213

In the ensuing sections, we will examine the methodology of our research, present our findings, and discuss their implications in the broader context of the interpretability of CNNs.

2. Literature Review

Over the previous decade, Convolutional Neural Networks have sustained unprecedented development and demonstrated transformative impacts on the realm of image processing. These deep learning models have consistently achieved state-of-the-art results in a diverse range of image-related tasks, from object detection to intricate segmentation. A comprehensive review of deep learning models for image segmentation demonstrated various techniques such as fully convolutional pixel-labeling networks, encoder-decoder architectures, and visual attention models [5]. With expanding surveys on the potential of these CNN models, more inventive methods of image segmentation that strive to achieve new state-of-the-art results and confront the challenges involving CNN interpretability arose.

Various methods that address the interpretability challenges of CNNs have been proposed in recent years. One noteworthy method is Semantic-Enhanced Image Clustering (SIC). This method utilizes a visual-language pre-training model to enhance image clustering. Through the process of mapping images to a semantic space and generating pseudo-labels based on the image-semantic relationships, SIC is able to prevail over the need for ground-truth semantic labels and create semantically significant clusters [3]. This approach accentuates the importance of employing interdisciplinary approaches to image processing. Additionally, this approach reveals more information on the internal intricacies of CNNs through semantic implications.

Another promising method in the field of image clustering is Deep Adaptive Clustering (DAC). Applying feature learning, DAC focuses on addressing the challenge of clustering images through transforming the clustering tasks into a binary pairwise-classification framework task. This approach takes advantage of the preexisting specialties of deep convolutional networks which demonstrated outstanding results. Further exploring the mechanics of the DAC approach, the label features generated by the deep convolutional network are processed to identify their similarities based on the cosine distance between these features. The beauty of this approach is that the previous components manifest as one-hot vectors which streamlines image clustering [4]. The single-stage ConvNet-based approach has exhibited state-of-the-art performance on various elite datasets, demonstrating the further potential that is expected to be discovered.

With an emphasis on efficiency and effectiveness, a novel approach that employs K-means and fuzzy K-means clustering algorithms has proven to achieve impressive segmentation results. The K-means algorithm is traditionally used for data partitioning. However, it has found its niche in image segmentation by grouping pixel values based on their similarities in properties: color, intensity, or texture [6]. This results in distinct regions within an image that share common visual attributes. The fuzzy K-means augments this by allowing pixels to have different degrees of membership in different clusters. This is particularly beneficial for cases where the images contain overlapping or gradient regions, offering a more detailed segmentation. In essence, this method emerges as a potent tool in image analysis by ingeniously combining the straightforward nature of K-means with the adaptability of fuzzy logic.

In light of the above, despite the rapid advances of CNNs that have revolutionized image processing, the question of their internal workings remains. However, pioneering methods like SIC, DAC, and K-means clustering are foreshadowing a future where the enigma of the internal mechanics of CNNs is demystified.

3. Methodology

For this experiment, we selected the Tiny ImageNet dataset which is renowned for its diversity as it contains 200 classes of images. Each class contains 500 training images, 50 validation images, and 50 test images. The images were uniformly presented in a 64x64 pixels resolution [2]. As the VGG16 architecture demands the inputs to be 224x224 pixels [1], the images of our dataset were resized to the appropriate resolution. To further diversify our dataset and enhance our model's capabilities, techniques such as random horizontal flipping, random rotation, and random cropping were used. Additionally, to comply with the requirements of VGG16, the images were normalized to the appropriate values for both mean and standard deviation values, [0.485, 0.456, 0.406] and [0.229, 0.224, 0.225] respectively.

The VGG16 model has upheld a reputation for yielding outstanding results in classification tasks and

ISSN 2616-5775 Vol. 6, Issue 12: 118-124, DOI: 10.25236/AJCIS.2023.061213

having powerful feature extraction capabilities. Enabled by transfer learning, we initialized our model with the weights pre-trained on the original dataset used to train the VGG16, a subset of ImageNet, and fine-tuned the model with the data of Tiny ImageNet. Furthermore, we altered the final classification layer of our model to accommodate 200 output units, which parallels the number of classes in Tiny ImageNet. In training, we utilized the SGD optimizer [10] with a learning rate of 0.001 and a momentum of 0.9. We applied the batch size to be 250 given our system capabilities and ran the training for 20 epochs.

Post-training, our focus shifted to feature extraction. Logically, we extracted the features using the penultimate layer, which precedes the final classification layer. From the theoretical perspective, this layer is likely to capture and provide the greatest patterns, details, and insights on each image. Hence, the clustering results produced with these feature vectors are likely to be that of the greatest significance.

The clustering algorithm employed in this study is the simple yet powerful K-means. This clustering algorithm utilizes minimizing the sum of the distances squared to form clusters. For the purpose of this research, the cluster number was set to 20 as it was more likely to produce more interpretable and meaningful groups. Additionally, to account for the high dimensionality of the vectors produced with concerns to the fundamental concepts of K-means clustering [11], setting 20 as total clusters assists in minimizing singleton clusters.

4. Results

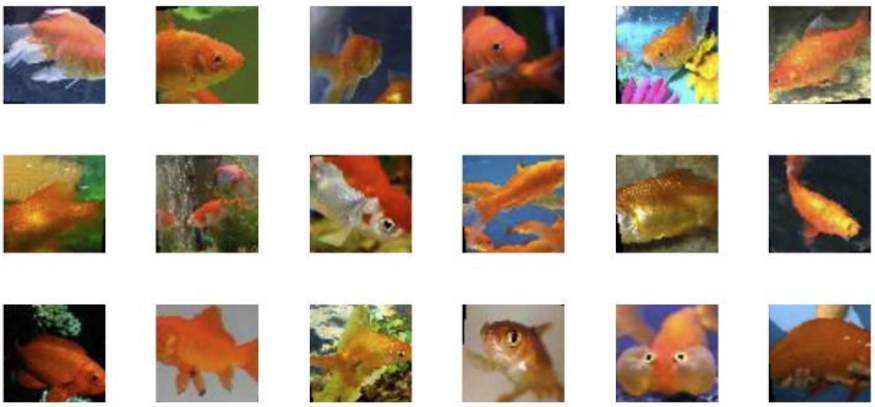
The output of the clustering is displayed in a manner such that 36 images per cluster are visible. The original output can be located at https://github.com/Wiliamz01/CNN-Driven-Clustering.

However, to highlight the intended overgeneralization employed and the interpretability of this study, the portions of clusters, which contain intriguing and relevant patterns, are presented in the following.

Cluster 0

Figure 1: A selection of images featuring animals and abaci from cluster 0 of results.

Cluster 2



ISSN 2616-5775 Vol. 6, Issue 12: 118-124, DOI: 10.25236/AJCIS.2023.061213

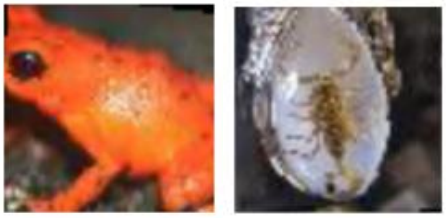


Figure 2: A selection of images presenting goldfish and other marine creatures of such colors from cluster 1 of results.

Cluster 3

Figure 3: A selection of images featuring crocodiles, fishes, and insects from cluster 3 of results.

Cluster 5

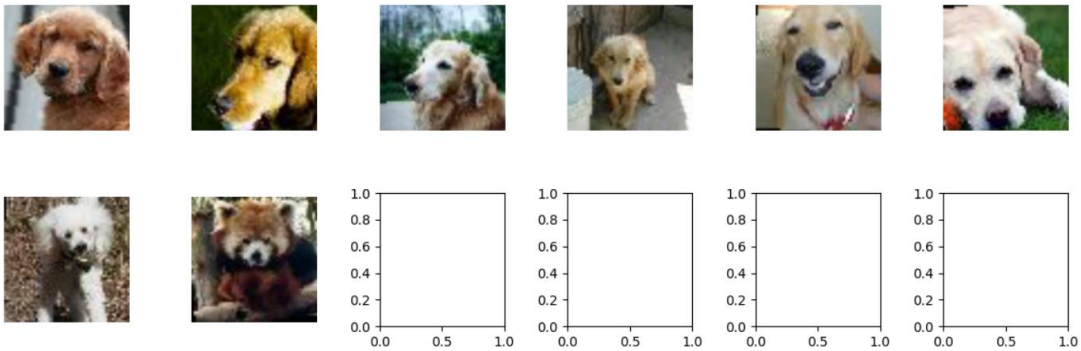


Figure 4: A selection of images featuring dogs from cluster 5 of results.

Cluster 6, 7, 10 respectively

ISSN 2616-5775 Vol. 6, Issue 12: 118-124, DOI: 10.25236/AJCIS.2023.061213

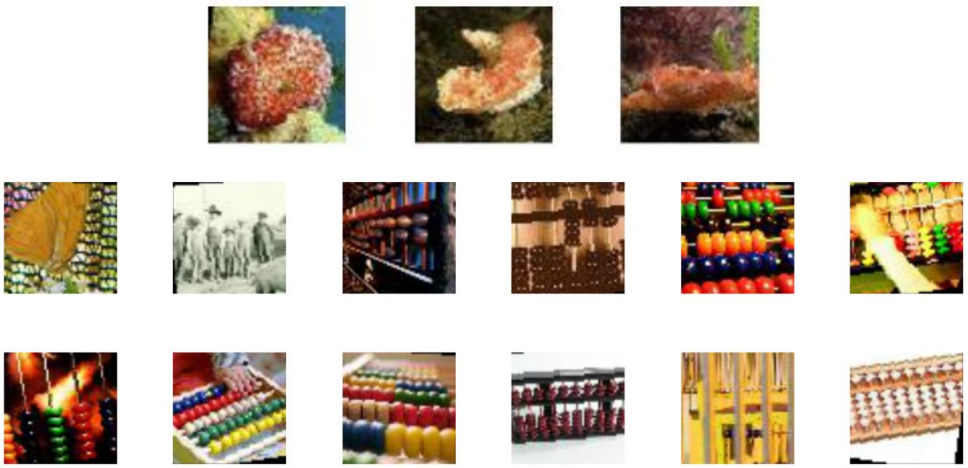


Figure 5: A selection of images from cluster 6, 7, and 10 representing various animals and abaci of color.

Cluster 10

Figure 6: A selection of images featuring humans and animals in a unique manner from cluster 10 of results.

Cluster 12

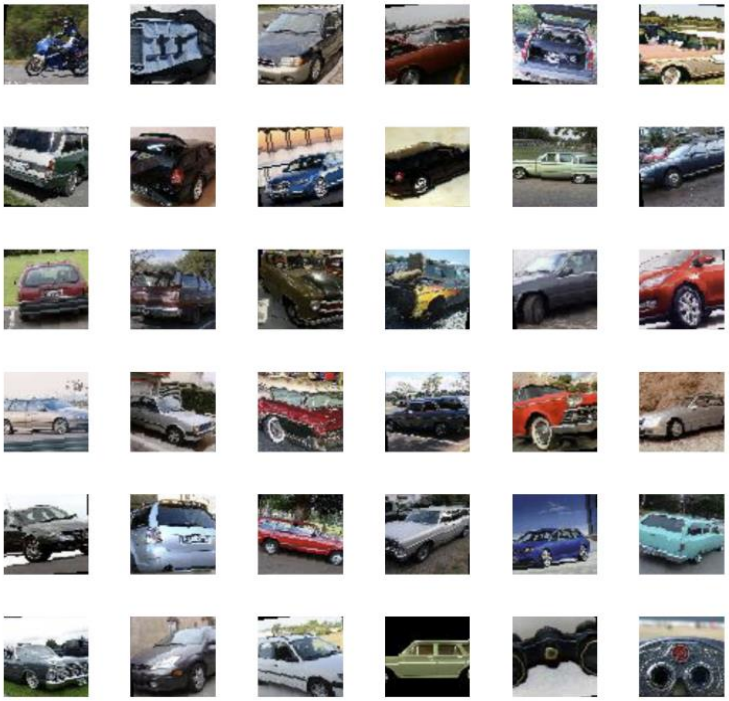


Figure 7: A selection of images presenting automobiles and mechanical parts from cluster 12 of results.

ISSN 2616-5775 Vol. 6, Issue 12: 118-124, DOI: 10.25236/AJCIS.2023.061213

Cluster 15

Figure 8: A brief selection of images representing humans with objects or collars surrounding their necks from cluster 15 of results.

5. Discussion

The results of our clustering experiment provide intriguing insights into the internal representations derived by the convolutional neural network. Our observation throughout the clusters inspected indicates several key aspects:

5.1. Feature Specificity

The clustering images with cages, especially those with humans near animals (Figure 1), demonstrates that the CNN has learned to recognize certain contextual features. This is further supported by the clustering of the images with abacuses, irrespective of their backgrounds. This pattern indicates that the model is capable of discerning intricate patterns and context within images.

5.2. Color and Shape Sensitivity

The cluster of goldfish and frogs (Figure 2) underscores the model's sensitivity to color. However, the inclusion of the frog, which morphologically is different from that of a fish, suggests that color plays a significant role in the interpretation of images in the model. However, the converse where shape is prioritized over color is also demonstrated in the clustering of crocodiles (Figure 3).

5.3. Orientation and Positional Awareness

The cluster that features the grouping of dogs based on their direction and angle of capture (Figure 4) indicates that the model is aware of object orientation and position. This implies that the CNNs can capture and prioritize spatial hierarchies within images.

5.4. Color Pattern Recognition

The model's ability to cluster images based on color patterns can be observed in the cluster with frogs, fish, and abacuses (Figure 5). This demonstrates the model's proficiency in pattern recognition. This is certainly a crucial aspect for a model to possess as it can enable it to distinguish objects with similar shapes but different patterns.

5.5. Contextual Understanding

The cluster containing images of people with seafood (Figure 6) was particularly fascinating to us. It suggests that the model might be picking up on more abstract or contextual relationships within the images. It implies that the model was able to learn beyond just colors and shapes.

5.6. Theme Detection

Through analyzing the cluster with automobile-based images (Figure 7), it is evident that the model was able to recognize the "wheel and structure" theme in the automobiles. This implies that the model has the capability of understanding broader themes within images to an adequate extent. The concept is further accentuated by the cluster where humans were presented with objects or collars around their neck regions (Figure 8).

It is evident that CNNs have evolved to the stage where they are capable of discerning not just primary features but also intricate patterns, contexts, and themes within images. For instance, the work on Deep Adaptive Image Clustering [3] underscores the importance of feature learning in image clustering. Our

ISSN 2616-5775 Vol. 6, Issue 12: 118-124, DOI: 10.25236/AJCIS.2023.061213

results show that when CCNs are trained effectively, they are capable of extracting highly nuanced features from images. This corresponds to why the results of Deep Adaptive Image Clustering were extremely outstanding.

However, the results yielded from our study raise questions on whether CNNs may contain inherent biases. For instance, the clustering that was influenced by color patterns may indicate a potential bias towards certain colors or patterns. This could lead to poor model performance for more diverse datasets.

In closing, our exploration into the internal representation of CNNs yielded valuable insights into their strengths and potential weaknesses. Although the goal directly ahead is to gain a comprehensive understanding of their internal workings, it is solely a step taken for the ultimate milestone of thoroughly exploring the great depths of the potential of convolutional Neural Networks.

6. Conclusions

The primary purpose of this study was to delve into the intricate internal representations of images by state-of-the-art convolutional neural networks. By employing K-Means clustering on the extracted features from these networks, we were able to discern the similarities and differences between images. Hence, light was shed on how CNNs represent images internally.

Our findings indicate that CNNs have the capability of capturing a wide array of features, ranging from color patterns to shapes and even abstract relations among images. These observations assist in unveiling the obscurities created by the “black box”. While existing methods such as Semantic-Enhanced Image Clustering (SIC) and Deep Adaptive Clustering (DAC) have contributed greatly to improving the interpretability of CNNs, our work contributes to this active area of research with empirical evidence on the internal feature representations of CNNs. Additionally, our work places a heavier emphasis on a high and comprehensive level while most other works delve into the nuances of CNNs.

This study paves a path for additional future research. For example, with incorporation of other clustering algorithms and CNN architecture could be experimented with to potentially produce more meaningful and promising results.

In summary, this study takes significant strides toward unveiling the internal workings of the CNNs. As the popularity and applications for these models continue to grow, understanding these internal mechanics will be crucial for discovering the full potential of CNNs and the broader realm of image processing.

References

[1] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition.” in International Conference of Learning Representations (ICLR), 9 2015.
[2] Le, Ya, and Xuan S. Yang. "Tiny ImageNet Visual Recognition Challenge." 2015. Activeloop. 1312.6034.pdf (arxiv.org)
[3] S. Cai, L. Qiu, X. Chen, Q. Zhang, and Longteng Chen. “Semantic-Enhanced Image Clustering.” arXiv:2208.09849v2, 2023, https://arxiv.org/pdf/2208.09849.pdf
[4]J. Chang, L. Wang, G. Meng, S. Xiang, and Chunhong Pan. “Deep Adaptive Image Clustering.” Proceedings of the International Conference on Computer Vision (ICCV), 2017, pp. 5879-5887.
[5]S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and Demetri Terzopoulos. “Image Segmentation Using Deep Learning: A Survey.” arXiv:2001.05566v5, 2020, https://arxiv.org/pdf/2001.05566.pdf
[6] V. K. Dehariya, S. K. Shrivastava and R. C. Jain, "Clustering of Image Data Set Using K-Means and Fuzzy K-Means Algorithms," 2010 International Conference on Computational Intelligence and Communication Networks, Bhopal, India, 2010, pp. 386-391, doi: 10.1109/CICN.2010.80.
[7] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps.” arXiv:1312.6034v2, 2014, https://arxiv.org/pdf/1312.6034.pdf
[8] A Mahendran and A. Vedaldi, “Visualizing deep convolutional neural networks using natural pre-images.” arXiv:1512.02017v3, 2016, https://arxiv.org/pdf/1512.02017.pdf
[9] A. Dosovitskiy and T. Brox, “Inverting Visual Representations with Convolutional Networks.”, Proceedings of the International Conference on Computer Vision (ICCV), 2016, pp. 4829-4837.
[10] S. Ruder, “An overview of gradient descent optimization algorithms.” arXiv:1609.04747v2, 2017, https://arxiv.org/pdf/1609.04747.pdf
[11] Jin, X., Han, J. (2011). K-Means Clustering. In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_425