Analyzing the Geometric Structure of CNN Embedding Spaces under Architectural and Regularization Variations

Eduardo Henrique Basilio de Carvalho
Departamento de Engenharia Eletrônica
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
eduardohbc@ufmg.br

***Abstract*—This paper investigates the relationship between the geometric structure of a Convolutional Neural Network's (CNN) embedding space and its classification performance on the CIFAR-10 dataset. We conduct two sets of experiments. The first set analyzes the impact of varying model architectures, from smaller, shallower networks to larger, deeper ones. The second set examines the effect of different dropout rates on a fixed architecture. We quantify the quality of the embedding space using four spatial metrics: intra-class distance, inter-class distance, Silhouette Score, and Fisher's Discriminant Ratio. Our findings reveal a strong correlation between these geometric metrics and the model's test-set accuracy. Specifically, the Silhouette Score shows a remarkably high correlation, suggesting that well-separated and compact class clusters in the embedding space are a strong indicator of a model's generalization capability. These results provide insights into how architectural design and regularization techniques like dropout shape the feature representations learned by CNNs.**

***Index Terms*—Convolutional Neural Networks, Embedding Space, Representation Learning, Deep Learning, CIFAR-10, Dropout, Model Architecture.**

I. Introduction

Convolutional Neural Networks (CNNs) have achieved state-of-the-art performance in a wide range of computer vision tasks. However, understanding the internal workings of these complex models remains a significant challenge. The penultimate layer of a CNN, often referred to as the embedding or feature space, provides a high-level representation of the input data that is then used for classification by the final layer. The geometric structure of this embedding space is believed to be crucial for the model's performance. A well-structured embedding space should exhibit high separability between different classes and high compactness within the same class.

This paper presents a systematic analysis of the relationship between the geometry of the embedding space and the classification accuracy of CNNs. We aim to answer the following questions:

1) How does the architecture of a CNN affect the geometric properties of its embedding space?
2) How does regularization, specifically dropout, influence these geometric properties?
3) Can we use spatial metrics of the embedding space to predict a model's generalization performance?

To address these questions, we conduct two sets of controlled experiments on the CIFAR-10 dataset. In the first experiment, we train several CNNs with varying architectures. In the second, we fix the architecture and vary the dropout rate. For each trained model, we extract the embeddings from the penultimate layer and compute a set of spatial metrics to quantify the quality of the embedding space. We then analyze the correlation between these metrics and the final test accuracy.

Our results demonstrate a strong and consistent correlation between the geometric quality of the embedding space and the model's performance. This suggests that these spatial metrics can serve as valuable tools for understanding and evaluating CNNs, potentially guiding model design and hyperparameter tuning.

II. Methodology

A. Dataset and Preprocessing

All experiments were conducted on the CIFAR-10 dataset [1], which consists of 60,000 32x32 color images in 10 classes. The dataset is split into 50,000 training images and 10,000 test images. We apply standard data augmentation techniques to the training set, including random cropping and horizontal flipping. Both training and test images are normalized using the mean and standard deviation of the CIFAR-10 dataset.

B. Base Model

Our base model is a VGG-style CNN. The architecture consists of a sequence of convolutional blocks followed by an adaptive average pooling layer and a final linear classifier. Each convolutional block contains two convolutional layers, each followed by a batch normalization layer and a ReLU activation function, and concludes with a max-pooling layer. This modular design allows us to easily modify the depth and width of the network. The embeddings for our analysis are extracted from the output of the adaptive average pooling layer, just before the final classification layer.

C. Spatial Metrics

We use four metrics to quantify the geometric structure of the embedding space:

1) **Intra-Class Distance:** The average Euclidean distance between each sample and the centroid of its class. A smaller value indicates tighter, more compact clusters.

$$d_{intra} = \frac{1}{C} \sum_{c=0}^{C-1} \left(\frac{1}{N_c} \sum_{i \in \text{class } c} ||x_i - \mu_c||_2 \right) \quad (1)$$

where C is the number of classes, N_c is the number of samples in class c, x_i is the embedding of sample i, and μ_c is the centroid of class c.

2) **Inter-Class Distance:** The average Euclidean distance between the centroids of all pairs of classes. A larger value indicates better separation between clusters.

$$d_{inter} = \frac{2}{C(C-1)} \sum_{i=0}^{C-1} \sum_{j=i+1}^{C-1} ||\mu_i - \mu_j||_2 \quad (2)$$

3) **Silhouette Score:** A composite measure of cluster cohesion and separation. The score ranges from -1 to 1, where a high value indicates that samples are well-matched to their own cluster and poorly matched to neighboring clusters.

4) **Fisher's Discriminant Ratio (FDR):** The ratio of the trace of the between-class scatter matrix (S_B) to the trace of the within-class scatter matrix (S_W). A higher FDR indicates better class separability.

$$FDR = \frac{\text{tr}(S_B)}{\text{tr}(S_W)} \quad (3)$$

D. Experimental Protocol

We conducted two main experiments.

1) Varying Model Architecture: In this experiment, we trained eight different CNN models with varying architectures, depths, and widths, as detailed in Table I. The learning rate and number of epochs were also varied to produce a range of test accuracies. For each model, we recorded the final test accuracy and computed the four spatial metrics on the embeddings of the entire training set.

2) Varying Dropout Rate: In this experiment, we used a fixed CNN architecture (two convolutional blocks with 64 channels in the final block) and varied the dropout rate applied before the final linear layer. We tested dropout rates of 0.0, 0.1, 0.25, and 0.5. For each dropout rate, we trained two models with different numbers of epochs to generate a spread of test accuracies. Similar to the first experiment, we recorded the test accuracy and computed the spatial metrics for each model.

III. Results and Discussion

A. Impact of Model Architecture

The results of the architecture experiment are summarized in Table I. As expected, larger models, such as 'Medium S', achieved higher test accuracies.

TABLE I
Results of the Architecture Experiment

Model	Accuracy (%)	dIntra	dInter	FDR	Silhou.
Tiny F	56.42	2.7156	3.7067	0.8404	0.0202
Tiny S	67.41	2.5487	3.3708	0.7988	0.0490
Small F	80.70	3.9643	6.3416	1.1778	0.1333
Small S	80.47	3.1121	4.8576	1.0971	0.1254
Medium F	82.93	4.8309	7.8238	1.2121	0.1362
Medium S	84.92	2.9335	4.8841	1.2528	0.1609
Wide Shlw	72.86	3.8825	4.8759	0.7145	0.0493
Deep Nrrw	83.95	4.3339	7.9408	1.4835	0.1852

Figure 1 shows the correlation between each spatial metric and the test accuracy. We observe a moderate positive correlation for inter-class distance and Fisher's Discriminant Ratio, as well as a strong positive correlation for the Silhouette Score. The latter has a Pearson correlation coefficient of 0.928 with test accuracy, indicating that it is a significant predictor of model performance. Intra-class distance shows a weaker, but still positive, correlation (0.578). This suggests that while cluster tightness is important, cluster separation is a more dominant factor for high accuracy.

B. Impact of Dropout

The results for the dropout experiment are shown in Table II. Here, increasing the dropout rate does not monotonically improve accuracy, as it depends on the number of training epochs. However, a moderate dropout of 0.25, when trained for enough epochs, yields the highest accuracy in this set of experiments.

TABLE II
Results of the Dropout Experiment

Model	Accuracy (%)	dIntra	dInter	FDR	Silhou.
0 F	60.07	2.6843	3.5957	0.8200	0.0334
0 S	68.42	2.8496	3.8952	0.8536	0.0523
1 F	59.17	2.6715	3.5695	0.8082	0.0344
1 S	66.40	2.7661	4.1017	1.0004	0.0592
25 F	65.88	2.4933	3.5595	0.9097	0.0481
25 S	68.81	2.4172	3.5003	0.9348	0.0584
5 F	63.19	2.2057	3.3311	1.0226	0.0393
5 S	66.51	2.1588	3.4752	1.2013	0.0503

The correlation plot for the dropout experiment (Figure 2) shows that most of the spatial metrics over the embedding space lost their correlation to accuracy. The Silhouette Score, however, remains a strong predictor with a correlation of 0.925. Interestingly, intra-class distance now shows a weak negative correlation (-0.092), suggesting that higher dropout rates, which are a form of regularization, tend to create tighter clusters, even if they don't always lead to higher accuracy.

IV. Conclusion

Our experiments demonstrate a clear and strong link between the geometric structure of a CNN's embedding space and its classification performance. Both architectural choices

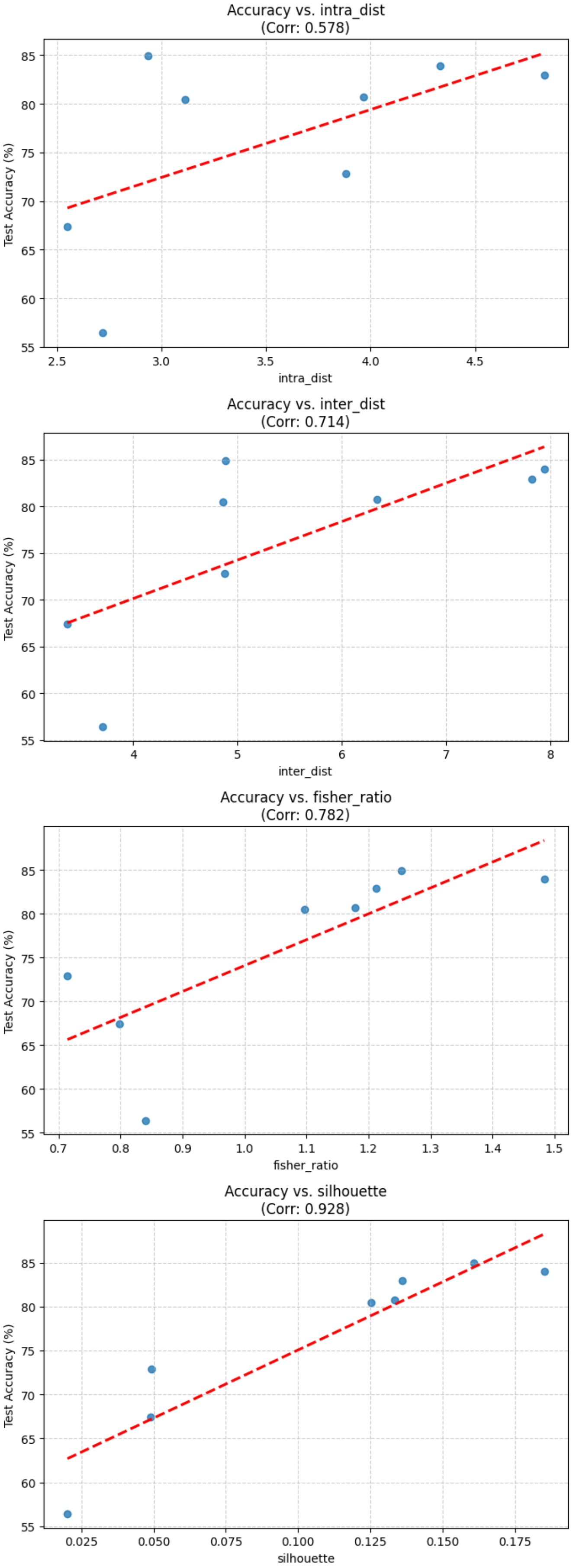

Fig. 1. Relationship between spatial metrics and test accuracy for varying architectures.

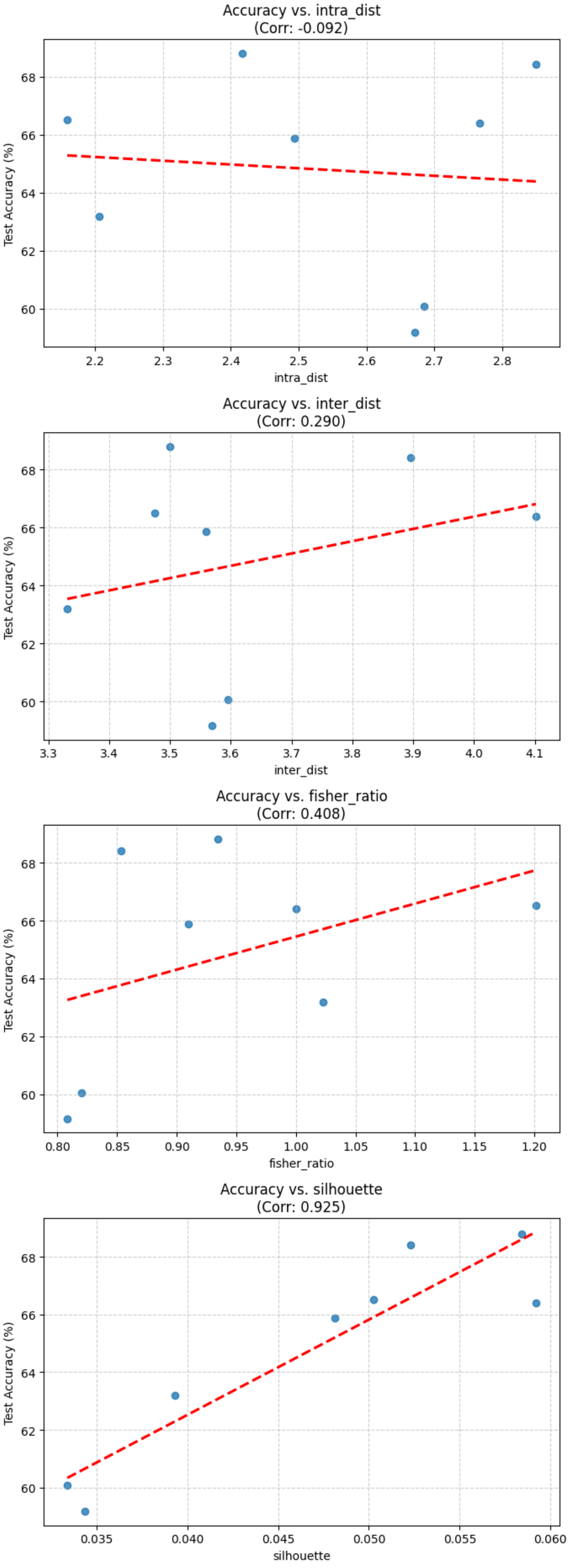

Fig. 2. Relationship between spatial metrics and test accuracy for varying dropout rates.

and regularization techniques like dropout significantly influence this structure. We found that Silhouette Score serves as an excellent predictor of a model's generalization ability. These findings highlight the importance of representation learning in deep neural networks and provide a quantitative framework for analyzing and comparing different models beyond just their final accuracy scores. Future work could extend this analysis to other datasets, architectures, and regularization methods, and explore the use of these metrics in automated model design and optimization.

References

[1] A. Krizhevsky, ”Learning Multiple Layers of Features from Tiny Images,” 2009.

[2] A. P. Braga, “Aprendendo com Exemplos: Princípios de Redes Neurais Artificiais e de Reconhecimento de Padrões,” Departamento de Engenharia Eletrônica, Escola de Engenharia, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil, Tech. Rep., 2025.