# Investigating Convolutional Neural Networks using Spatial Orderness

Rohan Ghosh
National University of Singapore
Singapore
rghosh92@gmail.com

Anupam K. Gupta
National University of Singapore
Singapore
anupamgupta1984@gmail.com

## Abstract

*Convolutional Neural Networks (CNN) have been pivotal to the success of many state-of-the-art classification problems, in a wide variety of domains (for e.g. vision, speech, graphs and medical imaging). A commonality within those domains is the presence of hierarchical, spatially agglomerative local-to-global interactions within the data. For instance in natural images, neighboring pixels are more likely contain similar values than non-neighboring pixels which are further apart. To that end, we propose a statistical metric called spatial orderness, which quantifies the extent to which the input data (2D) obeys the underlying spatial ordering at various scales. In our experiments, we mainly find that adding convolutional layers to a CNN could be counterproductive, when the data lacks spatial order at higher scales. Furthermore, we present a theoretical analysis (and empirical validation) of the spatial orderness of network weights, where we find that using smaller kernel sizes leads to kernels of greater spatial orderness and vice-versa.*

## 1. Introduction

There has been a large body of theoretical and experimental work exploring various attributes of CNNs which may contribute towards their excellent performance and generalization abilities ([14, 6, 1, 2]. However, the unusual effectivess of CNNs on a large variety of domains (vision, audio, graphs, medical imaging) is still not entirely comprehended.

Solely from the perspective of a mathematical function, it is intruiging to see a convolutional neural network demonstrate significant performance gains, when compared to a fully connected deep neural network. Empirical evidence exists [4, 10, 2] points to greater CNN depth being a key factor in better performance, but the same cannot be said of an FC-NN [11]. Unlike a FC-NN, CNN exhibits translation equivariance across its layers, that enable it to achieve translation equivariant representations deep within the network. This is useful for data containing global translational symmetries, but we see CNNs easily outperform FC-NNs in datasets such as MNIST where global translational symmetries do not exist [12].

Compared to structure-less FC-NNs, the inductive biases in a CNN are clearly better suited to handle classification problem in various domains. But, instead of a function-based introspection of a CNN, we ask which "convolution-conducive" characteristics of the data itself enable these inductive biases to flourish?

In this work, we systematically explore these questions, based on a hypothesis: *Convolutional structure in a neural network benefits from spatially ordered data.* We define *spatial order* to be the extent to which spatial proximity determines data value proximity. For images, an example of high spatial order in data is when spatially nearby pixels are more likely to have similar values than pixels which are far apart, and vice-versa. Spatially ordered data is likely to contain more meaningful spatial structure and hierarchy, and can benefit from locality-preserving feedforward functions; like convolutional operations in CNNs. [1]

A simple, novel metric for reliably quantifying spatial order in 2D data is proposed in this paper, denoted as *spatial orderness*. This metric can either be computed for a single 2D image, or for an entire dataset of 2D images. Theoretical results and extensive experiments (more in supplementary material) reveal how spatial orderness of the data is important to a CNN, and how it affects the spatial orderness of the kernels themselves.

## 2. Multi-Scale Spatial Orderness

First we define a spatial arrangement of pixel locations $(p, q, r)$ as a two-hop spatial arrangement, where $d(p, q) = d(q, r) = 1$ and $d(p, r) = 2$. Here $d(x, y)$ represents the distance in hops between pixel locations $x$ and $y$.

Given a set of images, $I_1, I_2, ..., I_k$, we first extract a fixed number ($l$) of triples of pixel intensities

[1] For graphs, we can extend the definition of spatial order to one of locality. For instance, a graph where every node is connected to every other node would be a counter-example of locality in a graph.

$(I_{n_1}(p_1), I_{n_1}(q_1), I_{n_1}(r_1)), .., (I_{n_l}(p_l), I_{n_l}(q_l), I_{n_l}(r_l))$, such that each triple of spatial locations $(p_i, q_i, r_i)$ follows a 2-hop spatial arrangement. Here $n_i$ denotes the image from which the $i^{th}$ triple was extracted. Next, we define the spatial orderness measure at the lowest scale as follows,

$$so(I)^1 = \left( \frac{\mathop{\mathbb{E}}_i \left[ \left( I_{n_i}(p_i) - I_{n_i}(r_i) \right)^2 \right]}{\mathop{\mathbb{E}}_i \left[ \left( I_{n_i}(p_i) - I_{n_i}(q_i) \right)^2 \right]} \right) - 1. \quad (1)$$

The metric is constrained to the range $(0, 1)$. With this, we can extend the definition of spatial orderness to multiple spatial scales. For that, a scale-space like decomposition is constructed by averaging $a \times a$ non-overlapping input regions onto a single pixel value. We let these sets of new mean downsampled images be denoted as $I_1^a, I_2^a, ...., I_k^a$. For each set of images at each scale, we denote their corresponding spatial orderness values by $so(I)^a$. At the end, we have a set of scalar values $so(I)^1, so(I)^2, ..., so(I)^p$, which represent the spatial orderness of the data at various scales. We summarize some the ways in which this measure can be interpreted:

- Randomly permuting the spatial locations of pixels (or blocks of pixels) will reduce spatial orderness to zero. Conversely, when a randomly permuted version of an input has an equal likelihood of occurence to its non-permuted form, the spatial orderness of data is zero at all scales.
- Spatial orderness at the lowest scale is indicative of how much more accurately the value of a pixel can be interpolated from a neighbor pixel *than* a non-neighbor pixel which is at a distance of 2 hops.

## 3. Experiments

The experiments reported herewith are conducted on the MNIST [8], Fashion-MNIST [13] and CIFAR-10 [7] datasets. For MNIST and Fashion-MNIST, we perform $2 \times 2$ max-pooling after each convolution layer, whereas for CIFAR-10, pooling was only performed after each alternate convolution layer. Additionally, a two layer fully connected network is used for CIFAR-10, whereas a single fc layer is used for MNIST and Fashion-MNIST. For consistency we used 64 units in all hidden layers, with kernels of size $3 \times 3$, except in section 4.1 (variation of kernel size).

### 3.1. Disrupting Spatial Orderness: Random Block-Swapping

Here we describe a method for disrupting the spatial orderness of the data at various scales, by performing block-swapping on the input. First we divide $(N \times N)$ images into blocks of size $k \times k$, such that $N/k \times N/k$ blocks span the entire image. Next, in each iteration of block swapping, a random chosen pair of image blocks are entirely swapped.

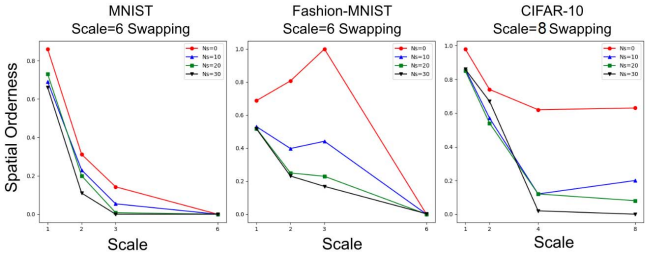


Figure 1. Spatial orderness of the MNIST, Fashion-MNIST and CIFAR-10 datasets at various scales, and their changes with block-swap operations performed on the data. For instance, the plots in red showcase the spatial orderness of the original, unswapped datasets ($Ns = 0$) at specific scales. For each dataset, block swapping was performed at a certain scale (specified on top).

We then repeat this process for $Ns$ number of iterations. More swaps (larger $Ns$) will lead to a greater disruption of spatial order, and thus should elicit lower values of spatial orderness, and vice-versa. Furthermore, the scale of the swap is relevant: swapping at a certain scale must not greatly impact the spatial orderness of lower scales, as the spatial arrangement in those scales is not overly affected. [2]

#### 3.1.1 Random Block-Swapping: Impact on Spatial Orderness

To analyze the effect of block-swapping on spatial orderness measures at various scales, we simply vary the number of block-swap operations on each image of the corresponding datasets. Increasing the number of swaps leads to a steady reduction of spatial order as a whole, in the data. Therefore, a metric which measures spatial order must return smaller values at the corresponding scale when the block swap operations on the image are increased. For our experiments, we block-swap at four sets of scales ($Ns = (0, 10, 20, 30)$) for the datasets of MNIST (at Scale=6), Fashion-MNIST (at Scale=6) and CIFAR-10 (at Scale=8), generating a total of 12 datasets: MNIST-swap$_6$(0,10,20,30), CIFAR10-swap$_8$(0,10,20,30) and Fashion-MNIST-swap$_6$(0,10,20,30). The results are shown in figure 1.

First, we expectedly observe that in all three datasets, spatial orderness at the highest scale is significantly lower than in the initial scales. This fact re-affirms the apparent "bag-of-features" like organisation of images at higher scales (objects or patterns are more positionally decorrelated at higher scales) (see [3]).

[2] Block-swapping at higher scales cannot altogether avoid disrupting the spatial order at lower scales, due to boundary effects of the blocks.

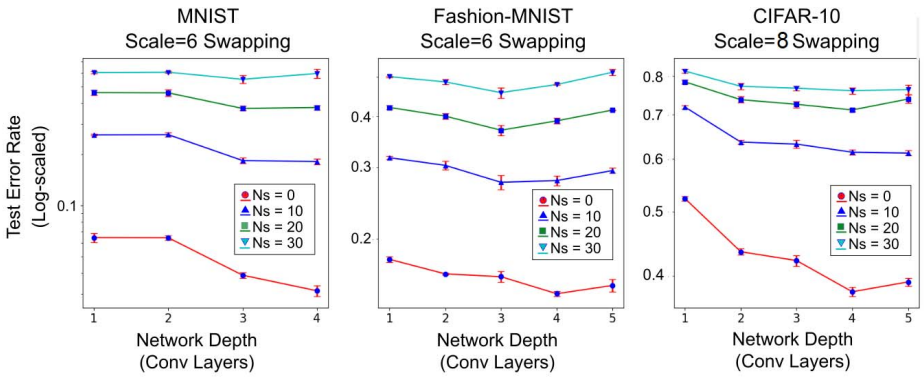


Figure 2. Semilog plots showing the test error rate of networks of different depths, trained on data corrupted by various degrees of spatial block-swapping ($Ns = (0, 10, 20, 30)$) on three different datasets (MNIST, Fashion-MNIST and CIFAR-10). Note that for data lacking in spatial order ($Ns > 0$), depth additions beyond a certain point do not yield improvements. Instead, such additions often significantly increase error rate, for larger $Ns$.

#### 3.1.2 Classification experiments: Is greater convolutional depth always better ?

Here we document CNN classification performance on MNIST-swap$_6$(0,10,20,30), CIFAR10-swap$_8$(0,10,20,30) and the Fashion-MNIST-swap$_6$(0,10,20,30) datasets. Our primary hypothesis is that convolution layers exploit the spatial orderness of data at multiple scales. Hence, for block-swapped data, we must expect the addition of convolution layers (beyond the scale of the swap) to pay decreasing dividends. Furthermore, because the block-swaps are only done at a higher scale, we should still find that adding initial convolution layers are beneficial, as spatial orderness of initial scales are still preserved (Figure 1).

Results are shown in figure 2. As hypothesized, we find that indeed adding convolution layers lead to decreasing gains, for larger number of block-swaps at the corresponding scales (larger $Ns$). Also, as anticipated, we observe that initial additions of convolution layers reduce test errors irrespective of block swapping. These findings are consistent with the theoretical results in [9].

## 4. Spatial Orderness of Kernels

### 4.1. Theoretical Results and Experimental Validation (Please see the Appendices)

We note that just like the inputs and the feature maps, one can treat the kernels (of size $K \times K$) as 2D images themselves. As such, it is also possible to compute the spatial orderness within the kernels, at the end of training. Convolution is linear in nature, and will elicit larger output responses when the input patches are highly correlated to the kernel form. Thus, kernels with very low spatial order (e.g. white noise kernels) are less likely to extract spatially structured and meaningful features, and vice-versa. Hence, from a feature extraction point of view, it is desirable that weights exhibit higher spatial orderness.

Here we summarize our theoretical results on the spatial orderness of kernels. Please find our main theoretical results (Theorems 1, Corollaries 1.1 and 1.2) and proofs in the supplementary material below. We summarize the theorems as follows.

- **Theorem 1 and Corollary 1.2**: **How is the spatial orderness of kernels and the spatial orderness of the feature map input related ?** We find that the spatial orderness of the kernels are likely to be higher when the inputs themselves have higher spatial orderness.
- **Corollary 1.1**: **How is the spatial orderness of kernels related to the choice of kernel size ?** We find that choosing a larger kernel size can lead to kernels with lower spatial orderness[3]. This shows that the choice of kernel size is quite important w.r.t ensuring spatially ordered kernels.

To verify the above theoretical results empirically, we train a CNN with 3 layers on a subset of MNIST. Figure 3 (a) and (b) shows the spatial orderness of kernels computed against variation of input spatial orderness, and kernel size respectively. We find that the experiments corroborate to our theoretical predictions.

These results add an interesting perspective on the debate of CNNs versus FC-NNs. Taken together, the results imply that a CNN is more likely to extract spatially ordered and meaningful features, subject to two necessary conditions: (a) the kernel size of the convolutions are small (i.e. more CNN than FC-NN like) and (b) the data on which the network is trained exhibits high spatial orderness.

[3]Note that by "spatial orderness of kernels" we mean the average spatial orderness of post-trained kernel weights (averaged across all kernels within a layer). In the following section, we empirically substantiate the results in the theorems.

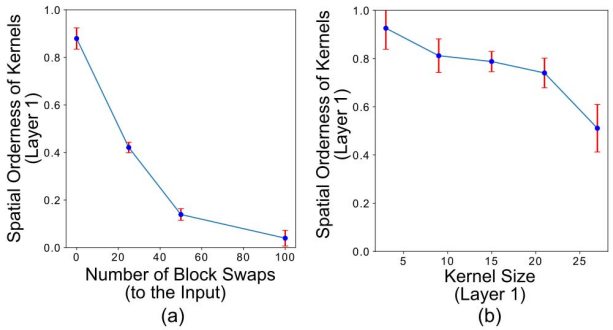


Figure 3. (a) demonstrates that disruption of spatial orderness at the input has an immediate effect on the spatial orderness of the kernels, and (b) shows that the size of kernels affect the spatial orderness of the trained kernels. All experiments were done on MNIST-1000 (1000 training examples used) and each experiment was repeated across six random splits of the data.

## 5. Discussions: Connection to Other Works

Recently it was found that on Imagenet, a bag-of-features based approach with shallow CNNs performs surprisingly close to bigger models which exploit spatial structure at higher scales [3]. Hence, spatial arrangement information beyond a certain scale is not very yielding in terms of improving classification performance. This is consistent with our findings in this paper. For instance, in section 3.1.1, it is observed that the spatial orderness of the image at higher scales is usually less than that of the lower scales, i.e. approaching a bag-of-features like organization.

Another example of testing the generalization abilities of CNNs is discussed in [5]. The authors observe that the CNN fails to generalize well when recognizability-preserving fourier domain filter masks were applied to the input. Throughout their experiments, the authors observe that the CNN trained on the low pass filtered radially-masked inputs showed the most consistent performance across datasets, having the smallest generalization gap. Our analysis on the spatial orderness of kernels in section 4 provides a possible explanation. Low-pass filtering enhances the spatial orderness of the input which ensures that trained kernels have greater spatial orderness; a reason for more consistent performance across data distortion variations.

## 6. Conclusions

A new statistical measure for quantifying spatial order within 2D data at various scales was proposed, called spatial orderness. This measure was shown to be indicative of the spatial organization at various scales, decreasing in value in correlation to the amount of input block-swapping performed. The performance gains from adding convolution layers was demonstrated to weaken with greater spatial order disruption. Theoretical and empirical results demonstrated the correlation between the spatial orderness of trained kernels, and the spatial orderness of the input. Additionally, we find that spatial orderness of kernels shows a significant drop with greater kernel-size, as it approaches a FC-NN like configuration.

## References


[1] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, Aug. 2013.

[2] Yoshua Bengio and Yann Lecun. *Scaling learning algorithms towards AI*. MIT Press, 2007.

[3] Wieland Brendel and Matthias Bethge. Approximating CNNs with bag-of-local-features models works surprisingly well on imagenet. In *International Conference on Learning Representations*, 2019.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

[5] Jason Jo and Yoshua Bengio. Measuring the tendency of cnns to learn surface statistical regularities. *ArXiv*, abs/1711.11561, 2017.

[6] Leslie Pack Kaelbling Kenji Kawaguchi and Yoshua Bengio. Generalization in deep learning. In *Mathematics of Deep Learning, Cambridge University Press, to appear. Prepint avaliable as: MIT-CSAIL-TR-2018-014, Massachusetts Institute of Technology*, 2018.

[7] Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 05 2012.

[8] Y. LECUN. The mnist database of handwritten digits. *http://yann.lecun.com/exdb/mnist/*.

[9] Hrushikesh Mhaskar and Tomaso A. Poggio. Deep vs. shallow networks : An approximation theory perspective. *CoRR*, abs/1608.03287, 2016.

[10] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[11] Shizhao Sun, Wei Chen, Liwei Wang, Xiaoguang Liu, and Tie-Yan Liu. On the depth of deep neural networks: A theoretical view. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 2066–2072. AAAI Press, 2016.

[12] Li Wan, Matthew Zeiler, Sixin Zhang, Yann LeCun, and Rob Fergus. Regularization of neural networks using dropconnect. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pages III–1058–III–1066. JMLR.org, 2013.

[13] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *ArXiv*, abs/1708.07747, 2017.

[14] Pan Zhou and Jiashi Feng. Understanding generalization and optimization performance of deep cnns. *International Conference on Machine Learning*, 2018.