Orthogonal Convolutional Neural Networks

Jiayun Wang Yubei Chen Rudrasis Chakraborty Stella X. Yu
UC Berkeley / ICSI
{peterwg,yubeic,rudra,stellayu}@berkeley.edu

Abstract

Deep convolutional neural networks are hindered by training instability and feature redundancy towards further performance improvement. A promising solution is to impose orthogonality on convolutional filters.

We develop an efficient approach to impose filter orthogonality on a convolutional layer based on the doubly block-Toeplitz matrix representation of the convolutional kernel instead of using the common kernel orthogonality approach, which we show is only necessary but not sufficient for ensuring orthogonal convolutions.

Our proposed orthogonal convolution requires no additional parameters and little computational overhead. This method consistently outperforms the kernel orthogonality alternative on a wide range of tasks such as image classification and inpainting under supervised, semi-supervised and unsupervised settings. Further, it learns more diverse and expressive features with better training stability, robustness, and generalization. Our code is publicly available.

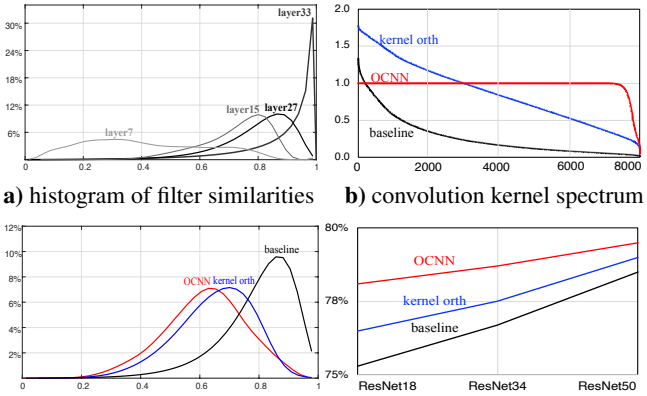
1. Introduction

While convolutional neural networks (CNNs) are widely successful [36, 14, 50], several challenges still exist: over parameterization or under utilization of model capacity [21, 12], exploding or vanishing gradients [7, 17], growth in saddle points [13], and shifts in feature statistics [31]. Through our analysis to solve these issues, we observe that convolutional filters learned in deeper layers are not only highly correlated and thus redundant (Fig.1a), but that each layer also has a long-tailed spectrum as a linear operator (Fig.1b), contributing to unstable training performance from exploding or vanishing gradients.

We propose *orthogonal CNN* (OCNN), where a convolutional layer is regularized with orthogonality constraints during training. When filters are learned to be as orthogonal as possible, they become de-correlated. Their filter responses are much less redundant. Therefore, the model capacity is better utilized, which improves the feature expressiveness and consequently the task performance.

Specifically, we show that simply by regularizing convolutions with our orthogonality loss during training, networks produce more uniform spectra (Fig.1b) and more diverse features (Fig.1c), delivering consistent performance gains with various network architectures (Fig.1d) on various tasks, e.g. image classification/retrieval, image inpainting, image generation, and adversarial attacks (Table 1).

Many works have proposed the orthogonality of linear operations as a type of regularization in training deep neural networks. Such a regularization improves the stability and



a) histogram of filter similarities **b)** convolution kernel spectrum

c) histogram of filter similarities **d)** classification accuracy

Figure 1. Our *OCNN* can remove correlations among filters and result in consistent performance gain over standard convolution *baseline* and alternative kernel orthogonality baseline (*kernel orth*) during testing. **a)** Normalized histograms of pairwise filter similarities of ResNet34 for ImageNet classification show increasing correlation among standard convolutional filters with depth. **b)** A standard convolutional layer has a long-tailed spectrum. While kernel orthogonality widens the spectrum, our OCNN can produce a more ideal uniform spectrum. **c)** Filter similarity (for layer 27 in **a**) is reduced most with our OCNN. **d)** Classification accuracy on CIFAR100 always increases the most with our OCNN.

Table 1. Summary of experiments and OCNN gains.

Task		Metric	Gain
Image Classification	CIFAR100	classification accuracy	3%
	ImageNet	classification accuracy	1%
	semi-supervised learning	classification accuracy	3%
Feature Quality	fine-grained image retrieval	kNN classification accuracy	3%
	unsupervised image inpainting	PSNR	4.3
	image generation	FID	1.3
	Cars196	NMI	1.2
Robustness	black box attack	attack time	7x less

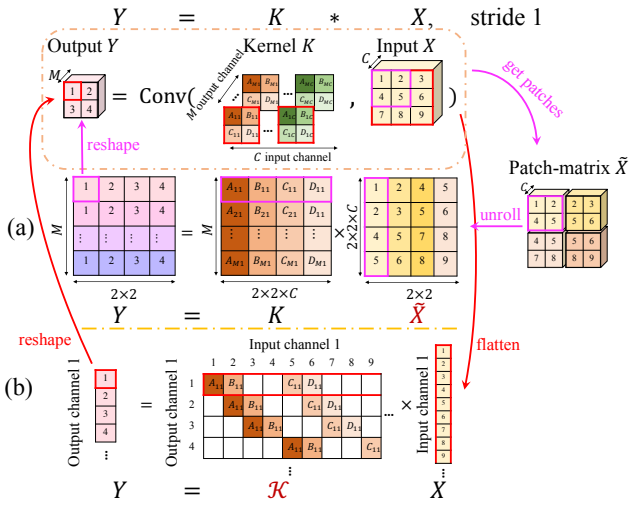


Figure 2. Basic idea of our OCNN. A convolutional layer $Y = \text{Conv}(K, X)$ can be formulated as matrix multiplications in two ways: **a)** *im2col* methods [58, 26] retain kernel K and convert input X to patch-matrix $\widetilde{X}$. **b)** We retain input X and convert K to a doubly block-Toeplitz matrix $\mathcal{K}$. With X and Y intact, we directly analyze the transformation from the input to the output. We further propose an efficient algorithm for regularizing $\mathcal{K}$ towards orthogonal convolutions and observe improved feature expressiveness, task performance and uniformity in $\mathcal{K}$'s spectrum (Fig.1b).

performance of CNNs [5, 57, 3, 4], since it can preserve energy, make spectra uniform [61], stabilize the activation distribution in different network layers [46], and remedy the exploding or vanishing gradient issues [1].

Existing works impose orthogonality constraints as kernel orthogonality, whereas ours directly implements orthogonal convolutions, based on an entirely different formulation of a convolutional layer as a linear operator.

Orthogonality for a convolutional layer $Y = \text{Conv}(K, X)$ can be introduced in two different forms (Fig.2).

1. **Kernel orthogonality** methods [57, 3, 4] view convolution as multiplication between the kernel matrix K and the *im2col* [58, 26] matrix $\widetilde{X}$, *i.e.* $Y = K\widetilde{X}$. The orthogonality is enforced by penalizing the disparity between the Gram matrix of kernel K and the identity matrix, *i.e.* $\|KK^T - I\|$. However, the construction of $\widetilde{X}$ from input X is also a linear operation $\widetilde{X} = QX$, and Q has a highly nonuniform spectrum.
2. **Orthogonal convolution** keeps the input X and the output Y intact by connecting them with a doubly block-Toeplitz (DBT) matrix $\mathcal{K}$ of filter K, i.e. $Y = \mathcal{K}X$ and enforces the orthogonality of $\mathcal{K}$ directly. We can thus directly analyze the linear transformation properties between the input X and the output Y.

Existing works on CNNs adopt kernel orthogonality, due to its direct filter representation.

We prove that kernel orthogonality is in fact only necessary but not sufficient for orthogonal convolutions. Consequently, the spectrum of a convolutional layer is still non-uniform and exhibits a wide variation even when the kernel matrix K itself is orthogonal (Fig.1b).

More recent works propose to improve the kernel orthogonality by normalizing spectral norms [40], regularizing mutual coherence [5], and penalizing off-diagonal elements [8]. Despite the improved stability and performance, the orthogonality of K is insufficient to make a linear convolutional layer orthogonal among its filters. In contrast, we adopt the DBT matrix form, and regularize $\|\text{Conv}(K, K) - I_r\|$ instead. While the kernel K is indirectly represented in the DBT matrix $\mathcal{K}$, the representation of input X and output Y is intact and thus the orthogonality property of their transformation can be directly enforced.

We show that our regularization enforces orthogonal convolutions more effectively than kernel orthogonality methods, and we further develop an efficient approach for our OCNN regularization.

To summarize, we make the following contributions.

1. We provide an equivalence condition for orthogonal convolutions and develop efficient algorithms to implement orthogonal convolutions for CNNs.
2. With no additional parameters and little computational overhead, our OCNN consistently outperforms other orthogonal regularizers on image classification, generation, retrieval, and inpainting under supervised, semi-supervised, and unsupervised settings.

Better feature expressiveness, reduced feature correlation, more uniform spectrum, and enhanced adversarial robustness may underlie our performance gain.

2. Related Works

***Im2col*-Based Convolutions**. The *im2col* method [58, 26] has been widely used in deep learning as it enables efficient GPU computation. It transforms the convolution into a General Matrix to Matrix Multiplication (GEMM) problem.

Fig.2a illustrates the procedure. **a)** Given an input X, we first construct a new input-patch-matrix $\widetilde{X} \in \mathbf{R}^{Ck^2 \times H'W'}$ by copying patches from the input and unrolling them into columns of this intermediate matrix. **b)** The kernel-patch-matrix $K \in \mathbf{R}^{M \times Ck^2}$ can then be constructed by reshaping the original kernel tensor. Here we use the same notation for simplicity. **c)** We can calculate the output $Y = K\widetilde{X}$ where we reshape Y back to the tensor of size $M \times H \times W$ – the desired output of the convolution.

The orthogonal kernel regularization enforces the kernel $K \in \mathbf{R}^{M \times Ck^2}$ to be orthogonal. Specifically, if $M \leq Ck^2$, the row orthogonal regularizer is $L_{\text{korth-row}} = \|KK^T - I\|_F$

where I is the identity matrix. Otherwise, column orthogonal may be achieved by $L_{\text{korth-col}} = \|K^T K - I\|_F$.

Kernel Orthogonality in Neural Networks. Orthogonal kernels help alleviate gradient vanishing or exploding problems in recurrent neural networks (RNNs) [15, 56, 10, 1, 54, 45]. The effect of soft versus hard orthogonal constraints on the performance of RNNs is discussed in [54]. A cheap orthogonal constraint based on a parameterization from exponential maps is proposed in [10].

Orthogonal kernels are also shown to stabilize the training of CNNs [46] and make more efficient optimizations [5]. Orthogonal weight initialization is proposed in [48, 39]; utilizing the norm-preserving property of orthogonal matrices, it is similar to the effect of batch normalization [31]. However, the orthogonality may not sustain as the training proceeds [48]. To ensure the orthogonality through the whole training, Stiefel manifold-based optimization methods are used in [22, 43, 30] and are further extended to convolutional layers in [43].

Recent works relax and extend the exact orthogonal weights in CNNs. Xie et al. enforce the Gram matrix of the weight matrix to be close to identity under Frobenius norm [57] . Bansal et al. further utilize mutual coherence and the restricted isometry property [5]. Orthogonal regularization has also been observed to help improve the performance of image generation in generative adversarial networks (GANs) [8, 9, 40].

All the aforementioned works adopt kernel orthogonality for convolutions. Sedghi et al. utilize the DBT matrix to analyze singular values of convolutional layers but do not consider orthogonality [49].

Feature Redundancy. Optimized CNNs are known to have significant redundancy between different filters and feature channels [32, 29]. Many works use the redundancy to compress or speed up networks [20, 25, 29]. The highly nonuniform spectra may contribute to the redundancy in CNNs. To overcome the redundancy by improving feature diversity, multi-attention [60], diversity loss [38], and orthogonality regularization [11] have been proposed.

Other Ways to Stabilize CNN Training. To address unstable gradient and co-variate shift problems, various methods have been proposed: Initialize each layer with near-constant variances [17, 23]; Use batch normalization to reduce internal covariate shifts [31]; Reparameterize the weight vectors and decouple their lengths from their directions [47]; Use layer normalization with the mean and variance computed from all of the summed inputs to the neurons [2]; Use a gradient norm clipping strategy to deal with exploding gradients and a soft constraint for vanishing gradients [45].

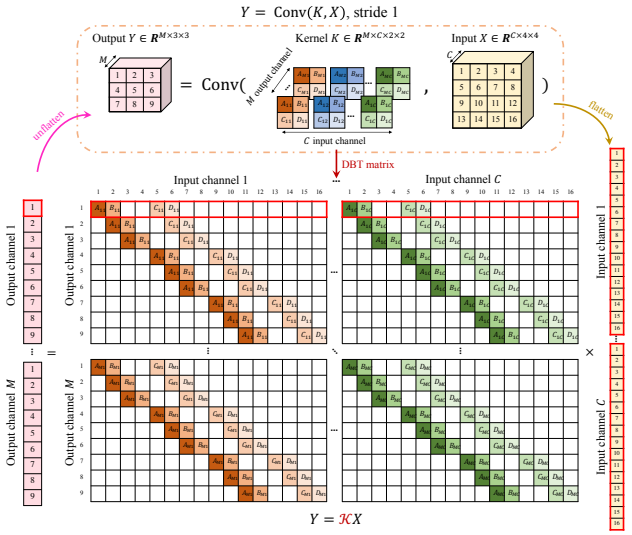


Figure 3. Convolution based on the doubly block-Toeplitz (DBT) matrix. We first flatten X to a vector $\mathbf{x}$, and then convert weight tensor $K \in \mathbf{R}^{M\times C\times k\times k}$ as DBT matrix $\mathcal{K} \in \mathbf{R}^{(MH'W')\times(CHW)}$. The output $\mathbf{y} = \mathcal{K}\mathbf{x}$. We can obtain the desired output $Y \in \mathbf{R}^{M\times H'\times W'}$ by reshaping $\mathbf{y}$. The example has input size $C \times 4 \times 4$, kernel size $M \times C \times 2 \times 2$ and stride 1.

3. Orthogonal Convolution

As we mentioned earlier, convolution can be viewed as an efficient matrix-vector multiplication, where matrix $\mathcal{K}$ is generated by a kernel K. In order to stabilize the spectrum of $\mathcal{K}$, we add convolutional orthogonality regularization to CNNs, which is a stronger condition than kernel orthogonality. First, we discuss the view of convolution as a matrix-vector multiplication in detail. Then, fast algorithms for constraining row and column orthogonality in convolutions are proposed. Condition 3 summarizes the orthogonality. In this work, we focus on the 2D convolution case, but concepts and conditions generalize to other cases.

3.1. Convolution as a Matrix-Vector Multiplication

For a convolutional layer with input tensor $X \in \mathbf{R}^{C\times H\times W}$ and kernel $K \in \mathbf{R}^{M\times C\times k\times k}$, we denote the convolution's output tensor $Y = \text{Conv}(K, X)$, where $Y \in \mathbf{R}^{M\times H'\times W'}$. We can further view K as M different filters, $\{K_i \in \mathbf{R}^{C\times k\times k}\}$. Since convolution is linear, we can rewrite $\text{Conv}(K, X)$ in a matrix-vector form:

$$Y = \text{Conv}(K, X) \Leftrightarrow \mathbf{y} = \mathcal{K}\mathbf{x} \tag{1}$$

where $\mathbf{x}$ is X flattened to a vector. Note that we adopt rigorous notations here while $\mathbf{x}$ and X are not distinguished previously. Each row of $\mathcal{K}$ has non-zero entries corresponding to a particular filter K_i at a particular spatial location. As a result, $\mathcal{K}$ can be constructed as a doubly block-Toeplitz (DBT) matrix $\mathcal{K} \in \mathbf{R}^{(MH'W')\times(CHW)}$ from kernel tensor $K \in \mathbf{R}^{M\times C\times k\times k}$.

We can obtain the output tensor Y by reshaping vector $\mathbf{y}$ back to the tensor form. Fig.3 depicts an example of a convolution based on DBT matrix, where we have input size of $C \times 4 \times 4$, kernel size of $M \times C \times 2 \times 2$ and stride 1.

3.2. Convolutional Orthogonality

Depending on the configuration of each layer, the corresponding matrix $\mathcal{K} \in \mathbf{R}^{(MH'W') \times (CHW)}$ may be a fat matrix ($MH'W' \leq CHW$) or a tall matrix ($MH'W' > CHW$). In either case, we want to regularize the spectrum of $\mathcal{K}$ to be uniform. In the fat matrix case, the uniform spectrum requires a row orthogonal convolution, while the tall matrix case requires a column orthogonal convolution, where $\mathcal{K}$ is a normalized frame [33] and preserves the norm.

In theory, we can implement the doubly block-Toeplitz matrix $\mathcal{K}$ and enforce the orthogonality condition in a brute force fashion. However, since $\mathcal{K}$ is highly structured and sparse, a much more efficient algorithm exists. In the following, we show the equivalent conditions to the row and column orthogonality, which can be easily computed.

Row Orthogonality. As we mentioned earlier, each row of $\mathcal{K}$ corresponds to a filter K_i at a particular spatial location (h', w') flattened to a vector, denoted as $\mathcal{K}_{ih'w',\bullet} \in \mathbf{R}^{CHW}$. The row orthogonality condition is:

$$\langle \mathcal{K}_{ih'_1w'_1,\bullet}, \mathcal{K}_{jh'_2w'_2,\bullet} \rangle = \begin{cases} 1, (i, h'_1, w'_1) = (j, h'_2, w'_2) \\ 0, \text{otherwise} \end{cases} \quad (2)$$

In practice, we do not need to check pairs when the corresponding filter patches do not overlap. It is clear that $\langle \mathcal{K}_{ih'_1w'_1,\bullet}, \mathcal{K}_{jh'_2w'_2,\bullet} \rangle = 0$ if either $|h_1 - h_2| \geq k$ or $|w_1 - w_2| \geq k$, since the two flattened vectors have no support overlap and thus have a zero inner product. Thus, we only need to check Condition 2 where $|h_1 - h_2|, |w_1 - w_2| < k$. Due to the spatial symmetry, we can choose fixed h_1, w_1 and only vary i, j, h_2, w_2, where $|h_1 - h_2|, |w_1 - w_2| < k$.

Fig.4 shows examples of regions of overlapping filter patches. For a convolution with the kernel size k and the stride S, the region to check orthogonality can be realized by the original convolution with padding $P = \lfloor \frac{k-1}{S} \rfloor \cdot S$. Now we have an equivalent condition to Condition 2 as the following self-convolution:

$$\text{Conv}(K, K, \text{padding} = P, \text{stride} = S) = I_{r0} \quad (3)$$

where $I_{r0} \in \mathbf{R}^{M \times M \times (2P/S+1) \times (2P/S+1)}$ is a tensor, which has zeros entries except for center $M \times M$ entries as an identity matrix. Minimizing the difference between $Z = \text{Conv}(K, K, \text{padding} = P, \text{stride} = S)$ and I_{r0} gives us a near row-orthogonal convolution in terms of DBT matrix $\mathcal{K}$.

Column Orthogonality. We use tensor $E_{i,h,w} \in \mathbf{R}^{C \times H \times W}$ to denote an input tensor, which has all zeros except a 1 entry at the i^{th} input channel, spatial location (h, w). Denoting $\mathbf{e_{ihw}} \in \mathbf{R}^{CHW}$ as the flattened vector of $E_{i,h,w}$, we can obtain a column $\mathcal{K}_{\bullet,ihw}$ of $\mathcal{K}$ by multiply $\mathcal{K}$ and vector $\mathbf{e_{ihw}}$:

$$\mathcal{K}_{\bullet,ihw} = \mathcal{K}\mathbf{e_{ihw}} = \text{Conv}(K, E_{i,h,w}) \quad (4)$$

Here, we slightly abuse the equality notation as the reshaping is easily understood. The column orthogonality condition is:

$$\langle \mathcal{K}_{\bullet,ih_1w_1}, \mathcal{K}_{\bullet,jh_2w_2} \rangle = \begin{cases} 1, (i, h_1, w_1) = (j, h_2, w_2) \\ 0, \text{otherwise} \end{cases} \quad (5)$$

Similar to the row orthogonality, since the spatial size of K is only k, Condition 5 only needs to be checked in a local region where there is spatial overlap between $\mathcal{K}_{\bullet,ih_1w_1}$ and $\mathcal{K}_{\bullet,jh_2w_2}$. For the stride 1 convolution case, there exists a simpler condition equivalent to Condition 5:

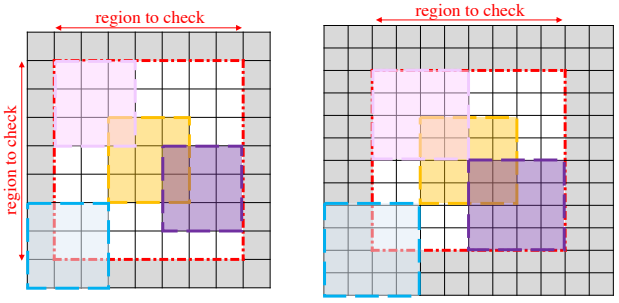
$$\text{Conv}(K^T, K^T, \text{padding} = k - 1, \text{stride} = 1) = I_{c0} \quad (6)$$

where K^T is the input-output transposed K, i.e. $K^T \in \mathbf{R}^{C \times M \times k \times k}$. $I_{c0} \in \mathbf{R}^{C \times C \times (2k-1) \times (2k-1)}$ has all zeros except for the center $C \times C$ entries as an identity matrix.

Comparison to Kernel Orthogonality. The kernel row- and column-orthogonality conditions can be written in the following convolution form respectively:

$$\begin{cases} \text{Conv}(K, K, \text{padding} = 0) = I_{r0} \\ \text{Conv}(K^T, K^T, \text{padding} = 0) = I_{c0} \end{cases} \quad (7)$$

where tensor $I_{r0} \in \mathbf{R}^{M \times M \times 1 \times 1}$, $I_{c0} \in \mathbf{R}^{C \times C \times 1 \times 1}$ are both equivalent to identity matrices[1].



(a) kernel size=3, stride=1 (b) kernel size=4, stride=2

Figure 4. The spatial region to check for row orthogonality. It is only necessary to check overlapping filter patches for the row orthogonality condition. We show two example cases: stride $S = 1$ with kernel size $k = 3$ and stride $S = 2$ with kernel size $k = 4$. In both examples, the orange patch is the center patch, and the red border is the region of overlapping patches. For example, pink and purple patches fall into the red region and overlap with the center region; blue patches are not fully inside the red region and they do not overlap with the orange ones. We can use padding to obtain the overlapping regions.

[1] Since there is only 1 spatial location.

Obviously, the kernel orthogonality conditions 7 are necessary but not sufficient conditions for the orthogonal convolution conditions 3,6 in general. For the special case when convolution stride is k, they are equivalent.

Row-Column Orthogonality Equivalence. The lemma below unifies the row orthogonality condition 2 and column orthogonality condition 5. This lemma [37] gives a uniform convolution orthogonality independent of the actual shape of $\mathcal{K}$ and provides a unique regularization: $\min_K L_{\text{orth}} = \|Z - I_{r0}\|_F^2$, which only depends on Condition 3.

Lemma 1. *The row orthogonality and column orthogonality are equivalent in the MSE sense, i.e.* $\|\mathcal{K}\mathcal{K}^T - I\|_F^2 = \|\mathcal{K}^T\mathcal{K} - I'\|_F^2 + U$*, where* U *is a constant.*

We leave the proof to Section D of supplementary materials.

Orthogonal Regularization in CNNs. We add an additional soft orthogonal convolution regularization loss to the final loss of CNNs, so that the task objective and orthogonality regularization can be simultaneously achieved. Denoting $\lambda > 0$ as the weight of the orthogonal regularization loss, the final loss is:

$$L = L_{\text{task}} + \lambda L_{\text{orth}} \tag{8}$$

where L_{task} is the task loss, e.g. softmax loss for image classification, and L_{orth} is the orthogonal regularization loss.

4. Experiments

We conduct 3 sets of experiments to evaluate OCNNs. The first set benchmarks our approach on image classification datasets CIFAR100 and ImageNet. The second set benchmarks the performance under semi-supervised settings and focuses on qualities of learned features. For high-level visual feature qualities, we experiment on the fine-grained bird image retrieval. For low-level visual features, we experiment on unsupervised image inpainting. Additionally, we compare visual feature qualities in image generation tasks. The third set of experiments focuses on the robustness of OCNN under adversarial attacks. We analyze OCNNs in terms of DBT matrix $\mathcal{K}$'s spectrum, feature similarity, hyperparameter tuning, and space/time complexity.

4.1. Classification on CIFAR100

The key novelty of our approach is the orthogonal regularization term on convolutional layers. We compare both conv-orthogonal and kernel-orthogonal regularizers on CIFAR-100 [35] and evaluate the image classification performance using ResNet [24] and WideResNet [59] as backbone networks. The kernel-orthogonality and our conv-orthogonality are added as additional regularization terms, without modifying the network architecture. Hence, the number of parameters of the network does not change.

ResNet and Row Orthogonality. Though we have derived a unified orthogonal convolution regularizer, we benchmark its effectiveness with two different settings. Convolutional layers in ResNet [24] usually preserve or reduce the dimension from input to output, *i.e.* a DBT matrix $\mathcal{K}$ would be a square or fat matrix. In this case, our regularizer leads to the row orthogonality condition. Table 2 shows top-1 classification accuracies on CIFAR100. Our approach achieves 78.1%, 78.7%, and 79.5% image classification accuracies with ResNet18, ResNet34 and ResNet50, respectively. For 3 backbone models, OCNNs outperform plain baselines by 3%, 2%, and 1%, as well as kernel orthogonal regularizers by 2%, 1%, and 1%.

WideResNet and Column Orthogonality. Unlike ResNet, WideResNet [59] has more channels and some tall DBT matrices $\mathcal{K}$. When the corresponding DBT matrix $\mathcal{K}$ of a convolutional layer increases dimensionality from the input to the output, our OCNN leads to the column orthogonality condition. Table 3 reports the performance of column orthogonal regularizers with backbone model of WideResNet28 on CIFAR100. Our OCNNs achieve 3% and 1% gain over plain baselines and kernel orthogonal regularizers.

Table 2. Top-1 accuracies on CIFAR100. Our OCNN outperforms baselines and the SOTA orthogonal regularizations.

	ResNet18	ResNet34	ResNet50
baseline [24]	75.3	76.7	78.5
kernel orthogonality [57]	76.5	77.5	78.8
OCNN (ours)	**78.1**	**78.7**	**79.5**

Table 3. WideResNet [59] performance. We observe improved performance of OCNNs.

	WideResNet [59]	Kernel orth [57]	OCNN
Acc.	77.0	79.3	**80.1**

4.2. Classification on ImageNet

We add conv-orthogonal regularizers to the backbone model ResNet34 on ImageNet [14], and compare OCNNs with state-of-the-art orthogonal regularization methods.

Experimental Settings. We follow the standard training and evaluation protocols of ResNet34. In particular, the total epoch of the training is 90. We start the learning rate at 0.1, decreasing by 0.1 every 30 epochs and weight decay 1e-4. The weight λ of the regularization loss is 0.01, the model is trained using SGD with momentum 0.9, and the batch size is 256.

Comparisons. Our method is compared with hard orthogonality OMDSM [30], kernel orthogonality [57] and spectral restricted isometry property regularization [5]. Table 4 shows the Top-1 and Top-5 accuracies on ImageNet. Without additional modification to the backbone model, OCNN achieves 25.87% top-5 and 7.89% top-1 error. The proposed method outperforms the plain baseline, as well as

other orthogonal regularizations by 1%.

Table 4. Top-1 and Top-5 errors on ImageNet [14] with ResNet34 [24]. Our conv-orthogonal regularization outperforms baselines and SOTA orthogonal regularizations.

	Top-1 error	Top-5 error
ResNet34 (baseline) [24]	26.70	8.58
OMDSM [30]	26.88	8.89
kernel orthogonality [57]	26.68	8.43
SRIP [5]	26.10	8.32
OCNN (ours)	**25.87**	**7.89**

4.3. Semi-Supervised Learning

A general regularizer should provide benefit to a variety of tasks. A common scenario that benefits from regularization is semi-supervised learning, where we have a large amount of data with limited labels. We randomly sample a subset of CIFAR100 as labeled and treat the rest as unlabeled. The orthogonal regularization is added to the baseline model ResNet18 without any additional modifications. The classification performance is evaluated on the entire validation set for all different labeled subsets.

We compare OCNN with kernel-orthogonal regularization while varying the proportion of labeled data from 10% to 80% of the entire dataset (Table 5). OCNN constantly outperforms the baseline by 2% - 3% under different fractions of labeled data.

Table 5. Top-1 accuracies on CIFAR100 with different fractions of labeled data. OCNNs are consistently better.

% of training data	10%	20%	40%	60%	80%	100%
ResNet18 [24]	31.2	47.9	60.9	66.6	69.1	75.3
kernel orthogonality [57]	33.7	50.5	63.0	68.8	70.9	76.5
Conv-orthogonality	**34.5**	**51.0**	**63.5**	**69.2**	**71.5**	**78.1**
Our gain	3.3	3.1	2.6	2.6	2.4	2.8

4.4. Fine-grained Image Retrieval

We conduct fine-grained image retrieval experiments on CUB-200 bird dataset [55] to understand high-level visual feature qualities of OCNNs. Specifically, we directly use the ResNet34 model trained on ImageNet (from Section 4.2) to obtain features of images in CUB-200, without further training on the dataset. We observed improved results with OCNNs (Fig.5). With conv-orthogonal regularizers, the top-1 k-nearest-neighbor classification accuracy improves from 25.1% to 27.0%, and top-5 k-nearest-neighbor classification accuracy improves from 39.4% to 42.3%.

4.5. Unsupervised Image Inpainting

To further assess the generalization capacity of OCNNs, we add the regularization term to the new task of unsupervised inpainting. In image inpainting, one is given an image X_0 with missing pixels in correspondence of a binary mask $M \in \{0,1\}^{C\times H\times W}$ of the same size of the image. The goal is to reconstruct the original image X by recovering missing pixels:

$$\min E(X; X_0) = \min \|(X - X_0) \odot M\|_F^2 \qquad (9)$$

Deep image prior (DIP) [52] proposed to use the prior implicitly captured by the choice of a particular generator network f_θ with parameter θ. Specifically, given a code vector/ tensor $\mathbf{z}$, DIP used CNNs as a parameterization $X = f_\theta(\mathbf{z})$. The reconstruction goal in Eqn.9 can be written as:

$$\min_\theta \|(f_\theta(\mathbf{z}) - X_0) \odot M\|_F^2 \qquad (10)$$

The network can be optimized without training data to recover X. We further add our conv-orthogonal regularization as an additional prior to the reconstruction goal, to validate if the proposed regularization helps the inpainting:

$$\min_\theta \|(f_\theta(\mathbf{z}) - X_0) \odot M\|_F^2 + \lambda L_{\text{orth}}(\theta) \qquad (11)$$

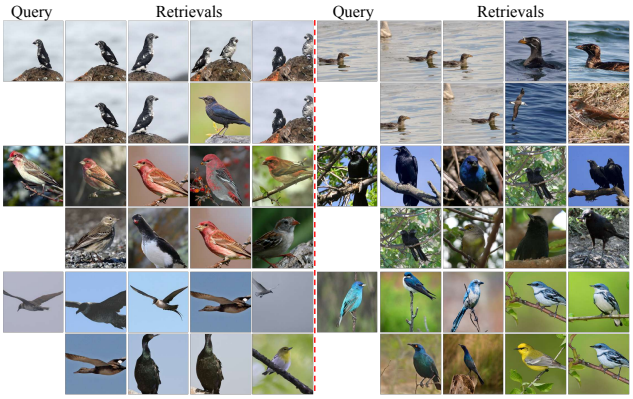


Figure 5. Image retrieval results on CUB-200 Birds Dataset. The model (ResNet34) is trained on ImageNet only. First row shows our OCNN results, while the second row shows the baseline model results. Ours achieves 2% and 3% top-1 and top-5 k-nearest neighbor classification gain.

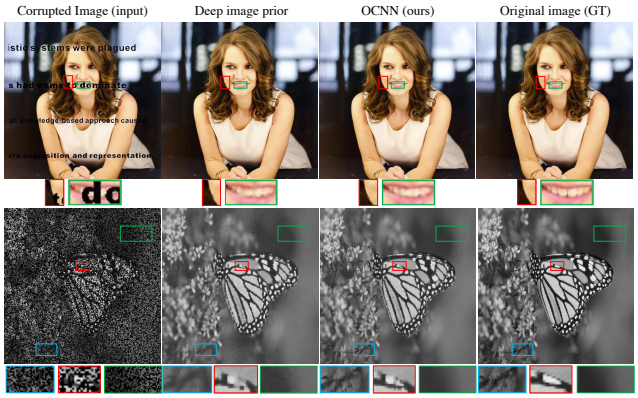


Figure 6. Image inpainting results compared with deep image prior [52]. Top – comparison on text inpainting example. Bottom – comparison on inpainting 50% of missing pixels. In both cases, our approach outperforms previous methods.

Table 6. Quantitative comparisons on the standard inpainting dataset [27]. Our conv-orthogonality outperforms the SOTA methods.

	Barbara	Boat	House	Lena	Peppers	C.man	Couple	Finger	Hill	Man	Montage
Convolutional dictionary learning [44]	28.14	31.44	34.58	35.04	31.11	27.90	31.18	31.34	32.35	31.92	28.05
Deep image prior (DIP) [52]	32.22	33.06	39.16	36.16	33.05	29.8	32.52	32.84	32.77	32.20	34.54
DIP + kernel orthogonality [57]	34.88	34.93	38.53	37.66	34.58	33.18	33.71	34.40	35.98	32.93	36.99
DIP + conv-orthogonality (ours)	**38.12**	**35.15**	**41.73**	**39.76**	**37.75**	**38.21**	**35.88**	**36.87**	**39.89**	**33.57**	**38.48**

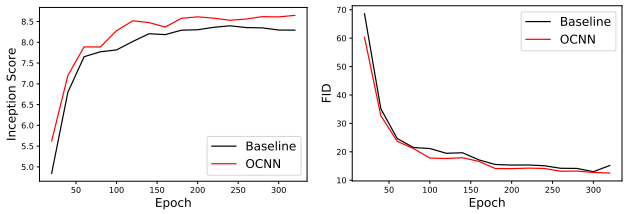


Figure 7. OCNNs have faster convergence for GANs. For IS (left) and FID (right), OCNNs consistently outperforms CNNs [18] at every epoch.

In the first example (Fig.6, top), the inpainting is used to remove text overlaid on an image. Compared with DIP [52], our orthogonal regularization leads to improved reconstruction result of details, especially for the smoothed face outline and finer teeth reconstruction.

The second example (Fig.6, bottom) considers inpainting with masks randomly sampled according to a binary Bernoulli distribution. Following the procedure in [44, 52], we sample a mask to randomly drop 50% of pixels. For a fair comparison, all the methods adopt the same mask. We observe improved background quality, as well as finer reconstruction of the texture of butterfly wings.

We report quantitative PSNR comparisons on the standard data set [26] in Table 6. OCNN outperforms previous state-of-the-art DIP [52] and convolutional sparse coding [44]. We also observe performance gains compared to kernel orthogonal regularizations.

4.6. Image Generation

Orthogonal regularizers have shown great success in improving the stability and performance of GANs [9, 40, 8]. We analyze the influence of convolutional orthogonal regularizers on GANs with the best architecture reported in [18]. Training takes 320 epochs with OCNN regularizer applied to both the generator and discriminator. The regularizer loss λ is set to 0.01 while other settings are retained as default.

The reported model is evaluated 5 times with 50k images each. We achieve an inception score (IS) of 8.63 ± 0.007 and Fréchet inception distance (FID) of 11.75 ± 0.04 (Table 7), outperforming the baseline and achieving the state-of-the-art performance. Additionally, we observe faster convergence of GANs with our regularizer (Fig.7).

4.7. Robustness under Attack

The uniform spectrum of $\mathcal{K}$ makes each convolutional layer approximately a 1-Lipschitz function. Given a small perturbation to the input, Δx, the change of the output Δy is bounded to be low. Therefore, the model enjoys robustness under attack. Our experiments demonstrate that it is much harder to search for adversarial examples.

Table 7. Inception Score and Fréchet Inception Distance comparison on CIFAR10. Our OCNN outperforms the baseline [18] by 0.3 IS and 1.3 FID.

	IS	FID
PixelCNN [53]	4.60	65.93
PixelIQN [42]	5.29	49.46
EBM [16]	6.78	38.20
SNGAN [40]	8.22	21.70
BigGAN [8]	**9.22**	14.73
AutoGAN [18]	8.32	13.01
OCNN (ours)	8.63	**11.75**

Table 8. Attack time and number of necessary attack queries needed for 90% successful attack rate.

	Attack time/s	# necessary attack queries
ResNet18 [24]	19.3	27k
OCNN (ours)	136.7	46k

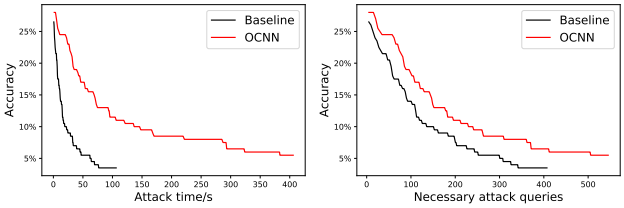


Figure 8. Model accuracy v.s. attack time and necessary attack queries. With our conv-orthogonal regularizer, it takes 7x time and 1.7x necessary attack queries to achieve 90% successful attack rate. Note that baseline ends at accuracy 3.5% while ours ends at 5.5% with the same iteration.

We adopt the simple black box attack [19] to evaluate the robustness of baseline and OCNN with ResNet18 [24] backbone architecture trained on CIFAR100. The attack samples around the input image and finds a "direction" to rapidly decrease the classification confidence of the network by manipulating the input. We only evaluate on the correctly classified test images. The maximum iteration is 10,000 with pixel attack. All other settings are retained. We report the attack time and number of necessary attack queries for a specific attack successful rate.

It takes approximately 7x time and 1.7x attack queries to attack OCNN, compared with the baseline (Table 8 and Fig.8). Additionally, after the same iterations of the attack, our model outperforms the baseline by 2%.

To achieve the same attack rate, baseline models need more necessary attack queries, and searching for such queries is nontrivial and time consuming. This may account for the longer attack time of the OCNN.

4.8. Analysis

To understand how the conv-orthogonal regularization help improve the performance of CNNs, we analyze several aspects of OCNNs. First, we analyze the spectrum of the DBT matrix $\mathcal{K}$ to understand how it helps relieve gradient vanishing/exploding. We then analyze the filter similarity of networks over different layers, followed by the influence of the weight λ of the regularization term. Finally, we analyze the time and space complexity of OCNNs.

Spectrum of the DBT Kernel Matrix $\mathcal{K}$. For a convolution $Y = \mathcal{K}X$, we analyze the spectrum of $\mathcal{K}$ to understand the properties of the convolution. We analyze the spectrum of $K \in \mathbf{R}^{64\times128\times3\times3}$ of the first convolutional layer of the third convolutional block of ResNet18 network trained on CIFAR100. For fast computation, we use input of size $64 \times 16 \times 16$, and solve all the singular values of $\mathcal{K}$.

As in Fig.1(b), the spectrum of plain models vanishes rapidly, and may cause gradient vanishing problems. Kernel orthogonality helps the spectrum decrease slower. With our conv-orthogonal regularization, the spectrum almost always stays at 1. The uniform spectrum perserves norm and information between convolutional layers.

Filter Similarity. Orthogonality makes off-diagonal elements become 0. This means that for any two channels of a convolutional layer, correlations should be relatively small. This can reduce the filter similarity and feature redundancy across different channels.

We use guided back-propagation patterns [51] on images from the validation set of the ImageNet [14] dataset to investigate filter similarities. Guided back-propagation patterns visualize the gradient of a particular neuron with respect to the input image $X \in \mathbf{R}^{C\times H\times W}$. Specifically for a layer of M channels, the combination of flattened guided back-propagation patterns is denoted as $G \in \mathbf{R}^{M\times CWH}$. The correlation matrix $\text{corr}(G)$ over different channels of this layer is $(\text{diag}(K_{GG}))^{-\frac{1}{2}} K_{GG} (\text{diag}(K_{GG}))^{-\frac{1}{2}}$, where $K_{GG} = \frac{1}{M}[(G - \mathbb{E}[G])(G - \mathbb{E}[G])^T]$ is the covariance matrix. We plot the histogram of off-diagonal elements of $\text{corr}(G)$ of all validation images.

Fig.1(a) depicts the normalized histogram of pairwise filter similarities of plain ResNet34. As the number of channels increases with depth from 128 to 512, the curve shifts right and becomes far narrower, i.e., more filters become similar. Fig.1(c) depicts the histogram of filter similarities at layer 27 of ResNet34 with different regularizers. OCNNs make the curve shift left and become wider, indicating it can enhance filter diversity and decrease feature redundancy.

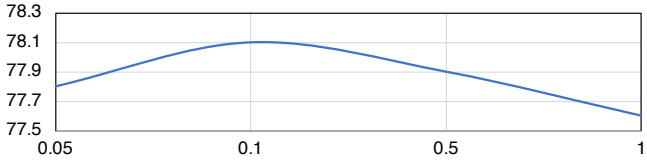


Figure 9. CIFAR100 classification accuracy (%) with different weight λ of the regularization loss. With backbone model ResNet18, we achieve the highest performance at $\lambda = 0.1$.

Hyper-Parameter Analysis. We analyze the influence of the weight λ of the orthogonality loss. As discussed earlier, we achieve the "soft orthogonality" by adding additional loss with weight λ to the network. Fig.9 plots the image classification performance of CIFAR100 with backbone model ResNet18 under λ ranging from 0.05 to 1.0. Our approach achieves the highest accuracy when $\lambda = 0.1$.

Space and Time Complexity. We analyze the space and time complexity in Table 9. The ResNet34 [24] backbone model is tested on ImageNet [14] with a single NVIDIA GeForce GTX 1080 Ti GPU and batch size 256.

The number of parameters and the test time of the CNN do not change since the regularizer is an additional loss term only used during training. With kernel orthogonal regularizers, the training time increases 3%; with conv-orthogonal regularizers, the training time increases 9%.

Table 9. Model size and training/ test time on ImageNet [14].

	ResNet34 [24]	kernel-orth [57]	OCNN
# Params.	21.8M	same	same
Training time (min/epoch)	49.5	51.0	54.1
Test time (min/epoch)	1.5	same	same

5. Summary

We develop an efficient OCNN approach to impose a filter orthogonality condition on a convolutional layer based on the doubly block-Toeplitz matrix representation of the convolutional kernel, as opposed to the commonly adopted kernel orthogonality approaches. We show that kernel orthogonality [5, 30] is necessary but not sufficient for ensuring orthogonal convolutions.

Our OCNN requires no additional parameters and little computational overhead, consistently outperforming the state-of-the-art alternatives on a wide range of tasks such as image classification and inpainting under supervised, semi-supervised and unsupervised settings. It learns more diverse and expressive features with better training stability, robustness, and generalization.

Acknowledgements. This research was supported, in part, by Berkeley Deep Drive, DARPA, and NSF-IIS-1718991. The authors thank Xudong Wang for discussions on filter similarity, Jesse Livezey for the pointer to a previous proof for row-column orthogonality equivalence, Haoran Guo, Ryan Zarcone, and Pratik Sachdeva for proofreading, and anonymous reviewers for their insightful comments.

References

[1] M. Arjovsky, A. Shah, and Y. Bengio. Unitary evolution recurrent neural networks. In *ICML*, pages 1120–1128, 2016. 2, 3

[2] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 3

[3] R. Balestriero and R. Baraniuk. Mad max: Affine spline insights into deep learning. *arXiv preprint arXiv:1805.06576*, 2018. 2

[4] R. Balestriero et al. A spline theory of deep networks. In *ICML*, pages 383–392, 2018. 2

[5] N. Bansal, X. Chen, and Z. Wang. Can we gain more from orthogonality regularizations in training deep cnns? In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4266–4276, 2018. 2, 3, 5, 6, 8

[6] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6541–6549, 2017. 11

[7] Y. Bengio, P. Simard, P. Frasconi, et al. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994. 1

[8] A. Brock, J. Donahue, and K. Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2019. 2, 3, 7

[9] A. Brock, T. Lim, J. M. Ritchie, and N. Weston. Neural photo editing with introspective adversarial networks. In *ICLR*, 2017. 3, 7

[10] M. L. Casado and D. Martínez-Rubio. Cheap orthogonal constraints in neural networks: A simple parametrization of the orthogonal and unitary group. In *ICML*, pages 3794–3803, 2019. 3

[11] Y. Chen, X. Jin, J. Feng, and S. Yan. Training group orthogonal neural networks with privileged information. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1532–1538, 2017. 3

[12] B. Cheung, A. Terekhov, Y. Chen, P. Agrawal, and B. A. Olshausen. Superposition of many models into one. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 10867–10876, 2019. 1

[13] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2933–2941, 2014. 1

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 1, 5, 6, 8, 11

[15] V. Dorobantu, P. A. Stromhaug, and J. Renteria. Dizzyrnn: Reparameterizing recurrent neural networks for norm-preserving backpropagation. *arXiv preprint arXiv:1612.04035*, 2016. 3

[16] Y. Du and I. Mordatch. Implicit generation and generalization in energy-based models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 7

[17] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 249–256, 2010. 1, 3

[18] X. Gong, S. Chang, Y. Jiang, and Z. Wang. Autogan: Neural architecture search for generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3224–3234, 2019. 7

[19] C. Guo, J. R. Gardner, Y. You, A. G. Wilson, and K. Q. Weinberger. Simple black-box adversarial attacks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2484–2493, 2019. 7

[20] S. Han, X. Liu, H. Mao, J. Pu, A. Pedram, M. A. Horowitz, and W. J. Dally. EIE: efficient inference engine on compressed deep neural network. In *IEEE Annual International Symposium on Computer Architecture (ISCA)*, pages 243–254, 2016. 3

[21] S. Han, H. Mao, and W. J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *ICLR*, 2016. 1

[22] M. Harandi and B. Fernando. Generalized backpropagation, étude de cas: Orthogonality. *arXiv preprint arXiv:1611.05927*, 2016. 3

[23] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, pages 1026–1034, 2015. 3

[24] 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 (CVPR)*, pages 770–778, 2016. 5, 6, 7, 8, 11

[25] Y. He, X. Zhang, and J. Sun. Channel pruning for accelerating very deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1389–1397, 2017. 3

[26] F. Heide, W. Heidrich, and G. Wetzstein. Fast and flexible convolutional sparse coding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5135–5143, 2015. 2, 7

[27] F. Heide, W. Heidrich, and G. Wetzstein. Fast and flexible convolutional sparse coding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 7

[28] E. Hoffer and N. Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer, 2015. 12

[29] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 3

[30] L. Huang, X. Liu, B. Lang, A. W. Yu, Y. Wang, and B. Li. Orthogonal weight normalization: Solution to optimization over multiple dependent stiefel manifolds in deep neural networks. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018. 3, 5, 6, 8

[31] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pages 448–456, 2015. 1, 3

[32] M. Jaderberg, A. Vedaldi, and A. Zisserman. Speeding up convolutional neural networks with low rank expansions. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2014. 3

[33] J. Kovačević, A. Chebira, et al. An introduction to frames. *Foundations and Trends in Signal Processing*, 2(1):1–94, 2008. 4

[34] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013. 12

[35] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, Canada, 2009. 5

[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012. 1

[37] Q. V. Le, A. Karpenko, J. Ngiam, and A. Y. Ng. Ica with reconstruction cost for efficient overcomplete feature learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1017–1025, 2011. 5, 12

[38] S. Li, S. Bak, P. Carr, and X. Wang. Diversity regularized spatiotemporal attention for video-based person reidentification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 369–378, 2018. 3

[39] D. Mishkin and J. Matas. All you need is a good init. In *ICLR*, 2016. 3

[40] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *ICLR*, 2018. 2, 3, 7

[41] Y. Movshovitz-Attias, A. Toshev, T. K. Leung, S. Ioffe, and S. Singh. No fuss distance metric learning using proxies. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 360–368, 2017. 12

[42] G. Ostrovski, W. Dabney, and R. Munos. Autoregressive quantile networks for generative modeling. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 3936–3945, 2018. 7

[43] M. Ozay and T. Okatani. Optimization on submanifolds of convolution kernels in cnns. *arXiv preprint arXiv:1610.07008*, 2016. 3

[44] V. Papyan, Y. Romano, J. Sulam, and M. Elad. Convolutional dictionary learning via local processing. In *ICCV*, 2017. 7

[45] R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. In *ICML*, pages 1310–1318, 2013. 3

[46] P. Rodríguez, J. Gonzàlez, G. Cucurull, J. M. Gonfaus, and F. X. Roca. Regularizing cnns with locally constrained decorrelations. In *ICLR*, 2017. 2, 3

[47] T. Salimans and D. P. Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 901–909, 2016. 3

[48] A. M. Saxe, J. L. McClelland, and S. Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *ICLR*, 2014. 3

[49] H. Sedghi, V. Gupta, and P. M. Long. The singular values of convolutional layers. In *ICLR*, 2019. 3

[50] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 1

[51] J. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR (workshop track)*, 2015. 8

[52] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 6, 7

[53] A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, et al. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4790–4798, 2016. 7

[54] E. Vorontsov, C. Trabelsi, S. Kadoury, and C. Pal. On orthogonality and learning recurrent networks with long term dependencies. In *ICML*, pages 3570–3578, 2017. 3

[55] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 6

[56] S. Wisdom, T. Powers, J. Hershey, J. Le Roux, and L. Atlas. Full-capacity unitary recurrent neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4880–4888, 2016. 3

[57] D. Xie, J. Xiong, and S. Pu. All you need is beyond a good init: Exploring better solution for training extremely deep convolutional neural networks with orthonormality and modulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6176–6185, 2017. 2, 3, 5, 6, 7, 8

[58] K. Yanai, R. Tanno, and K. Okamoto. Efficient mobile implementation of a cnn-based object recognition system. In *Proceedings of the International Conference on Multimedia (ICM)*, pages 362–366, 2016. 2

[59] S. Zagoruyko and N. Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2016. 5

[60] H. Zheng, J. Fu, T. Mei, and J. Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *ICCV*, pages 5209–5217, 2017. 3

[61] J. Zhou, M. N. Do, and J. Kovacevic. Special paraunitary matrices, cayley transform, and multidimensional orthogonal filter banks. *IEEE Transactions on Image Processing*, 15(2):511–519, 2006. 2

Supplementary Materials

The supplementary materials provide intuitive explanations of our approach (Section A), network dissection results to understand the change in feature redundancy/expressiveness (Section B), deep metric learning performance to understand the generalizability (Section C), proof of Lemma 1 (Section D) and visualizations of filter similarities (Section E).

A. Intuitive Explanations of our Approach

We analyze a convolutional layer which transforms input X to output Y with a learnable kernel K: $Y = \text{Conv}(K, X)$ in CNNs. Writing in linear matrix-vector multiplication form $Y = \mathcal{K}X$ (Fig.2b of the paper), we simplify the analysis from the perspective of linear systems. We do not use *im2col* form $Y = K\tilde{X}$ (Fig.2a of the paper) as there is an additional structured linear transform from X to $\tilde{X}$ and this additional linear transform makes the analysis indirect. As we mentioned earlier, the kernel orthogonality does not lead to a uniform spectrum.

The spectrum of $\mathcal{K}$ reflects the scaling property of the corresponding convolutional layer: different input X (such as cat, dog, and house images) would scale up by $\eta = \frac{\|Y\|}{\|X\|}$. The scaling factor η also reflects the scaling of the gradient. A typical CNN has highly non-uniform convolution spectrum (Fig.1b of the paper): for some inputs, it scales up to 2; for others, it scales by 0.1. For a deep network, these irregular spectra are multiplied together and can potentially lead to gradient exploding and vanishing issues.

Features learned by CNNs are also more redundant due to the non-uniform spectrum issues (Fig.1a of the paper). This comes from the diverse learning ability for different images and a uniform spectrum distribution could alleviate the problem. In order to achieve a uniform spectrum, We make convolutions orthogonal by enforcing DBT kernel matrix $\mathcal{K}$ to be orthogonal. The orthogonal convolution regularizer leads to uniform $\mathcal{K}$ spectra as expected. It further reduces feature redundancy and improves the performance (Fig.1b,c,d of the paper).

Besides classification performance improvements, we also observe improved visual feature qualities, both in high-level (image retrieval) and low-level (image inpainting) tasks. Our OCNNs also generate realistic images (Section 4.6) and are more robust to adversarial attacks (Section 4.7).

B. Network Dissection

We demonstrate in Section 4 that OCNNs reduce the feature redundancy by decorrelating different feature channels and enhancing the feature expressiveness with improved performance in image retrieval, inpainting, and generation. Network dissection [6] is utilized to further evaluate the feature expressiveness across different channels.

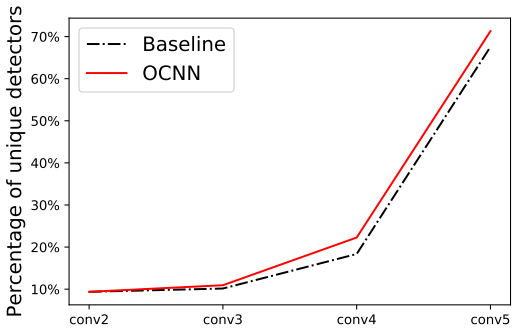


Figure 10. Percentage of unique detectors (mIoU ≥ 0.04) over different layers. Our OCNN has more unique detectors compared to plain baseline ResNet34 [24] at each layer.

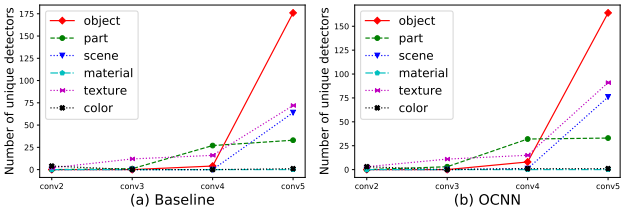


Figure 11. Distribution of concepts of unique detectors of different layers. Our OCNN has more uniform concept distribution compared to plain baseline ResNet34 [24].

Network dissection [6] is a framework that quantifies the interpretability of latent representations of CNNs by evaluating the alignment between individual hidden units and a set of semantic concepts. Specifically, we evaluate the baseline and our OCNN with backbone architecture ResNet34 [24] trained on ImageNet. Models are evaluated on the Broden [6] dataset, where each image is annotated with spatial regions of different concepts, including cat, dog, house, etc. The concepts are further grouped into 6 categories: scene, object, part, material, texture, and color. Network dissection framework compares the mean intersection over union (mIoU) between network channel-wise activation of each layer and ground-truth annotations. The units/feature channels are considered as "effective" when mIoU≥ 0.04. They are denoted as "unique detectors".

OCNNs have more unique detectors over different layers of the network (Table 10 & Fig.10). Additionally, OCNNs have more uniform distributions of 6 concept categories (Fig.11). These imply that orthogonal convolutions reduce feature redundancy and enhance the feature expressiveness.

Table 10. Number of unique detectors (feature channels with mIoU ≥ 0.04) comparisons on ImageNet [14].

	conv2	conv3	conv4	conv5
ResNet34 [24]	6	13	47	346
OCNN (ours)	6	14	57	365

C. Deep Metric Learning

We evaluate the generalizability and performance of our our orthogonal regularizer in deep metric learning tasks. Specifically, following the training/evaluation settings in [41], we perform retrieval and clustering on Cars196 dataset [34] and summarize the results in Table 11. We observe performance gains when the orthogonal regularizer is added.

Table 11. Retrieval/clustering performance on Cars196 (%).

	NMI	F1	Recall@1	@2	@4	@8
Triplet loss [28]	61.9	27.1	61.4	73.5	83.1	89.9
ProxyNCA [41]	62.4	29.2	67.9	78.2	85.6	90.6
[41]+Kernel orth	63.1	29.6	67.6	78.4	86.2	91.2
[41]+OCNN	**63.6**	**30.2**	**68.8**	**79.0**	**87.4**	**92.0**

D. Proof of the Orthogonality Equivalence

Here we provide a proof for the lemma 1: The row orthogonality and column orthogonality are equivalent in the MSE sense, i.e. $\|\mathcal{K}\mathcal{K}^T - I\|_F^2 = \|\mathcal{K}^T\mathcal{K} - I'\|_F^2 + U$, where U is a constant. A simple motivation for this proof is that when $\mathcal{K}$ is a square matrix, then $\mathcal{K}\mathcal{K}^T = I \iff \mathcal{K}^T\mathcal{K} = I'$. So we can hope to generalize this result and provide a more convenient algorithm. The following short proof is provided in the supplementary material of [37]. We would like to present it here for the reader's convenience.

Proof. It's sufficient to prove the general result, where we choose $\mathcal{K} \in \mathbf{R}^{M\times N}$ to be an arbitrary matrix[2]. We denote $\|\mathcal{K}\mathcal{K}^T - I_M\|^2$ as L_r and $\|\mathcal{K}^T\mathcal{K} - I_N\|^2$ as L_c.

$$
\begin{aligned}
L_r &= \|\mathcal{K}\mathcal{K}^T - I_M\|_F^2 \\
&= tr\left[(\mathcal{K}\mathcal{K}^T - I_M)^T(\mathcal{K}\mathcal{K}^T - I_M)\right] \\
&= tr(\mathcal{K}\mathcal{K}^T\mathcal{K}\mathcal{K}^T) - 2tr(\mathcal{K}\mathcal{K}^T) + tr(I_M) \\
&= tr(\mathcal{K}\mathcal{K}\mathcal{K}^T\mathcal{K}) - 2tr(\mathcal{K}^T\mathcal{K}) + tr(I_N) + M - N \\
&= tr\left[\mathcal{K}^T\mathcal{K}\mathcal{K}^T\mathcal{K} - 2\mathcal{K}^T\mathcal{K} + I_N\right] + M - N \\
&= tr\left[(\mathcal{K}^T\mathcal{K} - I_N)(\mathcal{K}^T\mathcal{K} - I_N)\right] + M - N \\
&= \|\mathcal{K}^T\mathcal{K} - I_N\|_F^2 + M - N \\
&= L_c + U
\end{aligned}
$$

where $U = M - N$.

□

E. Filter Similarity Visualizations

As shown in Fig.1, filter similarity increases with depth of the network. We visualize the guided back-propagation patterns to understand this phenomenon.

For the ResNet34 trained on ImageNet, we plot guided back-propagation patterns of an image in Fig.12. The first row depicts patterns of the first 3 channels from layer 7, while the second row depicts patterns of the first 3 channels from layer 33. Patterns of different channels from earlier layers are more diverse, while patterns of different channels from later layers usually focus on certain regions. The filter similarity increases with depth.

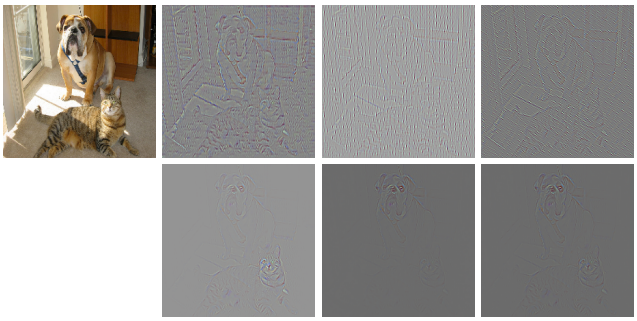


Figure 12. Guided back-propagation patterns of the input image (first column) with a ResNet34 model. The first row depicts patterns of the first 3 channels from layer 7, while the second row depicts patterns of the first 3 channels from layer 33. The filter similarity increases with network depth.

[2]Here M and N are just some constant, different from the the ones used in the main text.