Scaling the Scattering Transform: Deep Hybrid Networks

Edouard Oyallon
Département Informatique
Ecole Normale Supérieure
Paris, France
edouard.oyallon@ens.fr

Eugene Belilovsky
University of Paris-Saclay
INRIA and KU Leuven
eugene.belilovsky@inria.fr

Sergey Zagoruyko
Université Paris-Est
École des Ponts ParisTech
Paris, France
sergey.zagoruyko@enpc.fr

Abstract

We use the scattering network as a generic and fixed initialization of the first layers of a supervised hybrid deep network. We show that early layers do not necessarily need to be learned, providing the best results to-date with pre-defined representations while being competitive with Deep CNNs. Using a shallow cascade of 1×1 *convolutions, which encodes scattering coefficients that correspond to spatial windows of very small sizes, permits to obtain AlexNet accuracy on the imagenet ILSVRC2012. We demonstrate that this local encoding explicitly learns invariance w.r.t. rotations. Combining scattering networks with a modern ResNet, we achieve a single-crop top 5 error of* 11.4% *on imagenet ILSVRC2012, comparable to the Resnet-18 architecture, while utilizing only 10 layers. We also find that hybrid architectures can yield excellent performance in the small sample regime, exceeding their end-to-end counterparts, through their ability to incorporate geometrical priors. We demonstrate this on subsets of the CIFAR-10 dataset and on the STL-10 dataset.*

1. Introduction

Image classification is a high dimensional problem that requires building lower dimensional representations that reduce the non-informative images variabilities. For example, some of the main source of variability are often due to geometrical operations such as translations and rotations. An efficient classification pipeline necessarily builds invariants to these variabilities. Deep architectures build representations that lead to state-of-the-art results on image classification tasks [13]. These architectures are designed as very deep cascades of non-linear end-to-end learned modules [22]. When trained on large-scale datasets they have been shown to produce representations that are transferable to other datasets [42, 15], which indicate they have captured generic properties of a supervised task that consequently do not need to be learned. Indeed several works indicate geometrical structures in the filters of the earlier layers [19, 39] of Deep CNNs. However, understanding the precise operations performed by those early layers is a complicated [38, 26] and possibly intractable task. In this work we investigate if it is possible to replace these early layers, by simpler cascades of non-learned operators that reduce variability while retaining discriminative information.

Indeed, there can be several advantages to incorporating pre-defined geometric priors, via a hybrid approach of combining pre-defined and learned representations. First, end-to-end pipelines can be data hungry and ineffective when the number of samples is low. Secondly, it could permit to obtain more interpertable classification pipelines which are amenable to analysis. Finally, it can reduce the spatial dimensions and the required depth of the learned modules.

A potential candidate for an image representation is the SIFT descriptor [23] that was widely used before 2012 as a feature extractor in classification pipelines [30, 31]. This representation was typically encoded via an unsupervised Fisher Vector (FV) and fed to a linear SVM. However, several works indicate that this is not a generic enough representation to build further modules on top of [21, 2]. Indeed end-to-end learned features produce substantially better classification accuracy. A major improvement over SIFT can be found in the scattering transform [24, 6, 33], which is a type of deep convolutional network, which permits to retain discriminative information normally discarded by methods like SIFT while introducing geometric invariances and stability. Scattering transforms have been shown to already produce representations that lead to the top results on complex image datasets when compared to other unsupervised representations (even learned ones) [27]. This makes them an excellent candidate for the initial layers of a deep network. We thus investigate the use of scattering as a generic representation to combine with deep neural networks.

Related to our work [28] proposed a hybrid representation for large scale image recognition combining a prede-

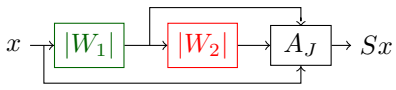


Figure 1. A scattering network. A_J concatenates the averaged signals.

fined representation and Neural Networks (NN), that uses Fisher Vector encoding of SIFT and leverages NNs as scalable classifiers. In contrast we use the scattering transform in combination with convolutional architectures. Our main contributions are as follows: First, we demonstrate that using supervised local descriptors, obtained by shallow 1×1 convolutions, with very small spatial window sizes permits to obtain AlexNet accuracy on the imagenet classification task (Subsection 2.3). We show empirically these encoders build explicit invariance to local rotations (Subsection 3.2). Second, we propose hybrid networks that combine scattering with modern CNNs (Section 4) and show that using scattering and a ResNet of reduced depth, we obtain similar accuracy to ResNet-18 on Imagenet (Subsection 4.1). Finally, we demonstrate in Subsection 4.3 that scattering permits a substantial improvement in accuracy in the setting of limited data.

Our highly efficient GPU implementation of the scattering transform is, to our knowledge, orders of magnitude faster than any other implementations, and allows training very deep networks applying scattering on the fly. Our scattering implementation [1] and pre-trained hybrid models [2]are available.

2. Scattering Networks and Hybrid Architectures

We introduce the scattering transform and motivate its use as a generic input for supervised tasks. A scattering network belongs to the class of CNNs whose filters are fixed as wavelets [27]. The construction of this network has strong mathematical foundations [24], meaning it is well understood, relies on few parameters and is stable to a large class of geometric transformations. In general, the parameters of this representation do not need to be adapted to the bias of the dataset [27], making its output a suitable generic representation.

We then propose and motivate the use of supervised CNNs built on top of the scattering network. Finally we propose a supervised encodings of scattering coefficients using 1x1 convolutions, that can retain interpertability and locality properties.

2.1. Scattering Networks

In this section, we recall the definition of the scattering transform. Consider a signal $x(u)$, with u the spatial position index and an integer $J \in \mathbb{N}$, which is the spatial scale of our scattering transform. Let ϕ_J be a local averaging filter with a spatial window of scale 2^J (here, a Gaussian smoothing function). Applying the local averaging operator, $A_J x(u) = x \star \phi_J(2^J u)$ we obtain the zeroth order scattering coefficient, $S_0 x(u) = A_J x(u)$. This operation builds an approximate invariant to translations smaller than 2^J, but it also results in a loss of high frequencies that are necessary to discriminate signals.

A solution to avoid the loss of high frequency information is provided by the use of wavelets. A wavelet is an integrable and localized function in the Fourier and space domain, with zero mean. A family of wavelets is obtained by dilating a complex mother wavelet ψ (here, a Morlet wavelet) such that $\psi_{j,\theta}(u) = \frac{1}{2^{2j}} \psi(r_{-\theta} \frac{u}{2^j})$, where $r_{-\theta}$ is the rotation by $-\theta$, and $j \geq 0$ is the scale of the wavelet. A given wavelet $\psi_{j,\theta}$ has thus its energy concentrated at a scale j, in the angular sector θ. Let $L \in \mathbb{N}$ be an integer parametrizing a discretization of $[0, 2\pi]$. A wavelet transform is the convolution of a signal with the family of wavelets introduced above, with an appropriate downsampling:

$$W_1 x(j_1, \theta_1, u) = \{x \star \psi_{j_1,\theta_1}(2^{j_1} u)\}_{j_1 \leq J, \theta_1 = 2\pi \frac{l}{L}, 1 \leq l \leq L}$$

Observe that j_1 and θ_1 have been discretized: the wavelet is chosen to be selective in angle and localized in Fourier. With appropriate discretization [27], $\{A_J x, W_1 x\}$ is approximatively an isometry on the set of signals with limited bandwidth, and this implies the energy of the signal is preserved. This operator then belongs to the category of multiresolution analysis operators, each filter being excited by a specific scale and angle, but with the output coefficients not being invariant to translation. To achieve invariance we can not apply A_J to $W_1 x$ since it gives a trivial invariant, namely zero.

To tackle this issue, we apply a non-linear point-wise complex modulus to $W_1 x$, followed by an averaging A_J, which builds a non trivial invariant. Here, the mother wavelet is analytic, thus $|W_1 x|$ is regular [1] which implies that the energy in Fourier of $|W_1 x|$ is more likely to be contained in a lower frequency domain than $W_1 x$. Thus, A_J preserves more energy of $|W_1 x|$. It is possible to define $S_1 x = A_J |W_1| x$, which can also be written as: $S_1 x(j_1, \theta_1, u) = |x \star \psi_{j_1,\theta_1}| \star \phi_J(2^J u)$; this is the first order scattering coefficients. Again, the use of the averaging builds an invariant to translation up to 2^J.

[1] http://github.com/edouardoyallon/pyscatwave
[2] http://github.com/edouardoyallon/scalingscattering

Once more, we apply a second wavelet transform W_2, with the same filters as W_1, on each channel. This permits the recovery of the high-frequency lost due to the averaging applied to the first order, leading to $S_2x = A_J|W_2||W_1|$, which can also be written as $S_2x(j_1, j_2, \theta_1, \theta_2, u) = |x \star \psi_{j_1,\theta_1}| \star \psi_{j_2,\theta_2}| \star \phi_J(2^J u)$. We only compute increasing paths, e.g. $j_1 < j_2$ because non-increasing paths have been shown to bear no energy [6]. We do not compute higher order scatterings, because their energy is negligible [6]. We call $Sx(u)$ the final scattering coefficient corresponding to the concatenation of the order 0, 1 and 2 scattering coefficients, intentionally omitting the path index of each representation. In the case of colored images, we apply independently a scattering transform to each RGB channel of the image, which means $Sx(u)$ has a size equal to $3 \times \left(1 + JL + \frac{1}{2}J(J-1)L^2\right)$, and the original image is down-sampled by a factor 2^J [6].

This representation is proved to linearize small deformations [24] of images, be non-expansive and almost complete [10, 5], which makes it an ideal input to a deep network algorithm, that can build invariants to this local variability via a first linear operator. We discuss it as an ideal initialization in the next subsection.

2.2. Cascading a supervised Deep architecture

We now motivate the use of a supervised architecture on top of a scattering network. Scattering transforms have yielded excellent numerical results [6] on datasets where the variabilities are completely known, such as MNIST or FERET. In these task, the problems encountered are linked to sample and geometric variance and handling these variances leads to solving these problems. However, in classification tasks on more complex image datasets, such variabilities are only partially known as there are also non geometrical intra-class variabilities. Although applying the scattering transform on datasets like CIFAR or Caltech leads to nearly state-of-the-art results in comparison to other unsupervised representations there is a large gap in performance when comparing to supervised representations [27]. CNNs fill in this gap, thus we consider the use of deep neural networks utilizing generic scattering representations in order to reduce more complex variabilities than geometric ones.

Recent works [25, 7, 17] have suggested that deep networks could build an approximation of the group of symmetries of a classification task and apply transformations along the orbits of this group, like convolutions. This group of symmetry corresponds to some of the non-informative intra class variabilities, which must be reduced by a supervised classifier. [25] motivates that to each layer corresponds an approximated Lie group of symmetry, and this approximation is progressive, in the sense that the dimension of these groups is increasing with depth. For instance, the main linear Lie group of symmetry of an image is the translation group, $\mathbb{R}^2$. In the case of a wavelet transform obtained by rotation of a mother wavelet, it is possible to recover a new subgroup of symmetry after a modulus non-linearity, the rotation SO_2, and the group of symmetry at this layer is the roto-translation group: $\mathbb{R}^2 \ltimes SO_2$. If no non-linearity was applied, a convolution along $\mathbb{R}^2 \ltimes SO_2$ would be equivalent to a spatial convolution. Discovering explicitly the next new and non-geometrical groups of symmetry is however a difficult task [17]; nonetheless, the roto-translation group seems to be a good initialization for the first layers. In this work, we investigate this hypothesis and avoid learning those well-known symmetries.

Thus, we consider two types of cascaded deep network on top of scattering. The first, referred to as the *Shared Local Encoder* (SLE), learns a supervised local encoding of the scattering coefficients. We motivate and describe the SLE in the next subsection as an intermediate representation between unsupervised local pipelines, widely used in computer vision prior to 2012, and modern supervised deep feature learning approaches. The second, referred to as a hybrid CNN, is a cascade of a scattering network and a standard CNN architecture, such as a ResNet [13]. In the sequel we empirically analyse hybrid CNNs, which permits to greatly reduce the spatial dimensions on which convolutions are learned and can reduce sample complexity.

2.3. Shared Local Encoder for Scattering Representations

We now discuss the spatial support of different approaches, in order to motivate our local encoder for scattering. In CNNs constructed for large scale image recognition, the representations at a specific spatial location and depth depend upon large parts of the initial input image and thus mixes global information. For example, at depth 2 of [19], the effective spatial support of the corresponding filter is already 32 pixels (out of 224). The specific representations derived from CNNs trained on large scale image recognition are often used as representations in other computer vision tasks or datasets [40, 42].

On the other hand prior to 2012 local encoding methods led to state of the art performance on large scale visual recognition tasks [30]. In these approaches local neighborhoods of an image were encoded using method such as SIFT descriptors [23], HOG [9], and wavelet transforms [32]. They were also often combined with an unsupervised encoding, such as sparse coding [4] or Fisher Vectors(FVs) [30]. Indeed, many works in classical image processing or classification [18, 4, 30, 28] suggests that the local encoding of an image permit to describe efficiently an image. Additionally for some algorithms that rely on local neighbourhoods, the use of local descriptors is essential [23]. Observe that a representation based on local non overlapping spatial neighborhood is simpler to analyze, as there is no ad-hoc

mixing of spatial information. Nevertheless, on large scale classification, this approach was surpassed by fully supervised learned methods [19].

We show that it is possible to apply, a similarly local, yet supervised encoding algorithm to a scattering transform, as suggested in the conclusion of [28]. First observe that at each spatial position u, a scattering coefficient $S(u)$ corresponds to a descriptor of a local neighborhood of spatial size 2^J. As explained in the first Subsection 2.1, each of our scattering coefficients are obtained using a stride of 2^J, which means the final representation can be interpreted as a non-overlapping concatenation of descriptors. Then, let f be a cascade of fully connected layers that we identically apply on each $Sx(u)$. Then f is a cascade of CNN operators with spatial support size 1×1, thus we write $fSx \triangleq \{f(Sx(u))\}_u$. In the sequel, we do not make any distinction between the 1×1 CNN operators and the operator acting on $Sx(u), \forall u$. We refer to f as a *Shared Local Encoder*. We note that similarly to Sx, fSx corresponds to non-overlapping encoded descriptors. To learn a supervised classifier on a large scale image recognition task, we cascade fully connected layers on top of the SLE.

Combined with a scattering network, the supervised SLE, has several advantages. Since the input corresponds to scattering coefficients, whose channels are structured, the first layer of f is as well structured. We further explain and investigate this first layer in Subsection 3.2. Unlike standard CNNs, there is no linear combinations of spatial neighborhoods of the different feature maps, thus the analysis of this network need only focus on the channel axis. Observe that if f was fed with raw images, for example in gray scale, it could not build any non-trivial operation except separating different level sets of these images.

In the next section, we investigate empirically this supervised SLE trained on the ILSVRC2012 dataset.

3. Local Encoding of Scattering

We evaluate the supervised SLE on the Imagenet ILSVRC2012 dataset. This is a large and challenging natural color image dataset consisting of 1.2 million training images and $50,000$ validation images, divided into 1000 classes. We then show some unique properties of this network and evaluate its features on a separate task.

3.1. Shared Local Encoder on Imagenet

We first describe our training pipeline, which is similar to [41]. We trained our network for 90 epochs to minimize the standard cross entropy loss, using SGD with momentum 0.9 and a batch size of 256. We used a weight decay of 1×10^{-4}. The initial learning rate is 0.1, and is dropped off by 0.1 at epochs $30, 50, 70, 80$. During the training process, each image is randomly rescaled, cropped, and flipped as in [13]. The final crop size is 224×224. At testing, we rescale the image to a size of 256, and extract a center crop of size 224×224.

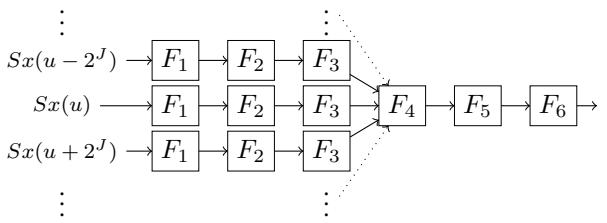


Figure 2. Architecture of the SLE, which is a cascade of 3 1×1 convolutions followed by 3 fully connected layers. The ReLU non-linearity are included inside the F_i blocks for clarity.

Method	Top 1	Top 5
FV + FC [28]	55.6	78.4
FV + SVM [30]	54.3	74.3
AlexNet	56.9	**80.1**
Scat + SLE	**57.0**	79.6

Table 1. Top 1 and Top 5 percentage accuracy reported from one single crop on ILSVRC2012. We compare to other local encoding methods, and SLE outperforms them. [28] single-crop result was provided by private communication.

We use an architecture which consists of a cascade of a scattering network, a SLE f, followed by fully connected layers. Figure 2 describes our architecture. We select the parameter $J = 4$ for our scattering network, which means the output representation has size $\frac{224}{2^4} \times \frac{224}{2^4} = 14 \times 14$ spatially and 1251 in the channel dimension. f is implemented as 3 layers of 1x1 convolutions F_1, F_2, F_3 with layer size 1024. There are 2 fully connected layers of ouput size 1524. For all learned layers we use batch normalization [16] followed by a ReLU [19] non-linearity. We compute the mean and variance of the scattering coefficients on the whole Imagenet, and standardized each spatial scattering coefficients with it.

Table 3.1 reports our numerical accuracies obtained with a single crop at testing, compared with local encoding methods, and the AlexNet that was the state-of-the-art approach in 2012. We obtain 20.4% at Top 5 and 43.0% Top 1 errors. The performance is analogous to the AlexNet [19]. In term of architecture, our hybrid model is analogous, and comparable to that of [30, 28], for which SIFT features are extracted followed by FV [31] encoding. Observe the FV is an unsupervised encoding compared to our supervised encoding. Two approaches are then used: either the spatial localization is handled either by a Spatial Pyramid Pooling [20], which is then fed to a linear SVM, either the spatial variables are directly encoded in the FVs, and classified with a stack of four fully connected layers. This last method is a major difference with ours, as the obtained de-

scriptor does not have a spatial indexing anymore which are instead quantified. Furthermore, in both case, the SIFT are densely extracted which correspond to approximatively $2\,10^4$ descriptors, whereas in our case, only $14^2 = 196$ scattering coefficients are extracted. Indeed, we tackle the non-linear aliasing (due to the fact the scattering transform is not oversampled) via random cropping during training, allowing to build an invariant to small translations. In Top 1, [30] and [28] obtain respectively 44.4% and 45.7%. Our method brings a substantial improvement of 1.4% and 2.7% respectively.

The BVLC AlexNet [3] obtains a of 43.1% single-crop Top 1 error, which is nearly equivalent to the 43.0% of our SLE network. The AlexNet has 8 learned layers and as explained before, large receptive fields. On the contrary, our training pipeline consists in 6 learned layers with constant receptive field of size 16×16, except for the fully connected layers that build a representation mixing spatial information from different locations. This is a surprising result, as it seems to suggest context information is only necessary at the very last layers, to reach AlexNet accuracy.

We study briefly the local SLE, which has only a spatial extent of 16×16, as a generic local image descriptor. We use the Caltech-101 benchmark which is a dataset of 9144 image and 102 classes. We followed the standard protocol for evaluation [4] with 10 folds and evaluate per class accuracy, with 30 training samples per class, using a linear SVM used with the SLE descriptors. Applying our raw scattering network leads to an accuracy of 62.8 ± 0.7, and the outputs features from F_1, F_2, F_3 brings respectively an absolute improvement of $13.7, 17.3, 20.1$. The accuracy of the final SLE descriptor is thus 82.9 ± 0.4, similar to that reported for the final AlexNet final layer in [42] and sparse coding with SIFT [4]. However in both cases spatial variability is removed, either by Spatial Pyramid Pooling [20], or the cascade of large filters. By contrasts the concatenation of SLE descriptors are completely local.

3.2. Interprating SLE's first layer

Finding structure in the kernel of the layers of depth less than 2 [39, 42] is a complex task, and few empirical analyses exist that shed light on the structure [17] of deeper layers. A scattering transform with scale J can be interpreted as a CNN with depth J [27], whose channels indexes correspond to different scattering frequency indexes, which is a structuration. This structure is consequently inherited by the first layer F_1 of our SLE f. We analyse F_1 and show that it builds explicitly invariance to local rotations, yet also that the Fourier bases associated to rotation are a natural bases of our operator. It is a promising direction to understand the nature of the two next layers.

[3] https://github.com/BVLC/caffe/wiki/Models-accuracy-on-ImageNet-2012-val

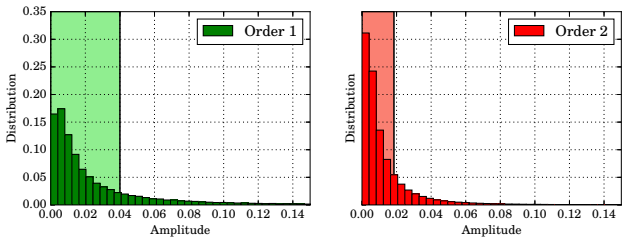


Figure 3. Histogram of $\hat{F}_1$ amplitude for first and second order coefficients. The vertical lines indicate a threshold that is used in Subsection 3.2 to sparsify $\hat{F}_1$. Best viewed in color.

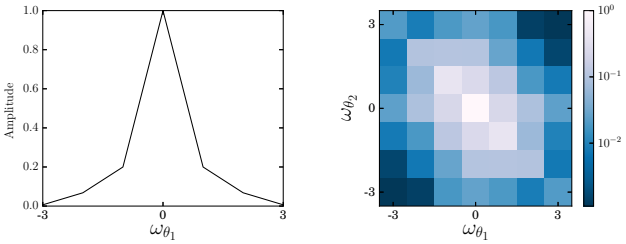


Figure 4. Energy $\Omega_1\{F\}$ (left) and $\Omega_2\{F\}$ (right) from Eq. 1 for given angular frequencies. Best viewed in color.

We first establish some mathematical notions linked to the rotation group that we use in our analysis. For the sake of clarity, we do not consider the roto-translation group. For a given input image x, let $r_\theta.x(u) \triangleq x(r_{-\theta}(u))$ be the image rotated by angle θ, which corresponds to the linear action of rotation on images. Observe the scattering representation is covariant with the rotation in the following sense:

$$\begin{aligned} S_1(r_\theta.x)(\theta_1, u) &= S_1x(\theta_1 - \theta, r_{-\theta}u) \triangleq r_\theta.(S_1x)(\theta_1, u) \\ S_2(r_\theta.x)(\theta_1, \theta_2, u) &= S_2x(\theta_1 - \theta, \theta_2 - \theta, r_{-\theta}u) \\ &\triangleq r_\theta.(S_2x)(\theta_1, \theta_2, u) \end{aligned}$$

Besides, in the case of the second order coefficients, (θ_1, θ_2) is covariant with rotations, but $\theta_2 - \theta_1$ is an invariant to rotation that correspond to a relative rotation.

Unitary representation framework [36] permits the building of a Fourier transform on compact group, like rotations. It is even possible to build a scattering transform on the roto-translation group [33]. Fourier analysis permits the measurement of the smoothness of the operator and, in the case of CNN operator, it is a natural basis.

We can now numerically analyse the nature of the operations performed along angle variables by the first layer F_1 of f, with output size $K = 1024$. Let us define as $\{F_1^0 S_0 x, F_1^1 S_1 x, F_1^2 S_2 x\}$ the restrictions of F_1 to the order 0,1,2 scattering coefficients respectively. Let $1 \leq k \leq K$ an index of a feature channel and $1 \leq c \leq 3$ the color index. In this case, $F_1^0 S_0 x$ is simply the weights associated to the smoothing $S_0 x$. $F_1^1 S_1 x$ depends only (k, c, j_1, θ_1), and F_1^2

depends on $(k, c, j_1, j_2, \theta_1, \theta_2)$. We would like to characterize the smoothness of these operators with respect to the variables (θ_1, θ_2), because Sx is covariant to rotations.

To this end, we define by $\hat{F}_1^1, \hat{F}_1^2$ the Fourier transform of these operators along the variables θ_1 and (θ_1, θ_2) respectively. These operator are expressed in the tensorial Frequency domain, which corresponds to a change of basis. In this experiment, we normalized each filter of F such that they have a ℓ_2 norm equal to 1, and each order of the scattering coefficients are normalized as well. Figure 3 shows the distribution of the amplitude of $\hat{F}_1^1, \hat{F}_2^2$. We observe that the distribution is shaped as a Laplace distribution, which is an indicator of sparsity.

To illustrate that this is a natural basis we explicitly sparsify this operator in its frequency basis and verify that empirically the network accuracy is minimally changed. We do this by thresholding by ϵ the coefficients of the operators in the Fourier domain. Specifically we replace the operators $\hat{F}_1^1$, $\hat{F}_1^2$ by $1_{|\hat{F}_1^1|>\epsilon}\hat{F}_1^1$ and $1_{|\hat{F}_1^2|>\epsilon}\hat{F}_1^2$. We select an ϵ that sets 80% of the coefficients to 0, which is indicated on Figure 3. *Without retraining* our network performance degrades by only an absolute value of 2% worse on Top 1 and Top 5 ILSVRC2012. We have thus shown that this basis permits a sparse approximation of the first layer, F_1. We now show evidence that this operator builds an explicit invariant to local rotations.

To aid our analysis we introduce the following quantities:

$$\Omega_1\{F\}(\omega_1) \triangleq \sum_{k,j_1,c} |\hat{F}_1^1(k, c, j_1, \omega_{\theta_1})|^2 \qquad (1)$$

$$\Omega_2\{F\}(\omega_{\theta_1}, \omega_{\theta_2}) \triangleq \sum_{k,c,j_1,j_2} |\hat{F}_1^2(k, c, j_1, j_2, \omega_{\theta_1}, \omega_{\theta_2})|^2$$

They correspond to the energy propagated by F_1 for a given frequency, and permit to quantify the smoothness of our first layer operator w.r.t. the angular variables. Figure 4 shows variation of $\Omega_1\{F\}$ and $\Omega_2\{F\}$ along frequencies. For example, if F_1^1 and F_1^2 were convolutional along θ_1 and (θ_1, θ_2), these quantities would correspond to their respective singular values. One sees that the energy is concentrated in the low frequency domain, which indicates that F_1 builds explicitly an invariant to local rotations.

4. Numerical performances of hybrid networks

We now demonstrate cascading modern CNN architectures on top of the scattering network can produce high performance classification systems. We apply hybrid convolutional networks on the Imagenet ILSVRC 2012 dataset as well as the CIFAR-10 dataset and show that they can achieve performance comparable to modern end-to-end learned approaches. We then evaluate the hybrid networks in the setting of limited data by utilizing a subset of CIFAR-10 as well as the STL-10 dataset and show that we can obtain substantial improvement in performance over analogous end-to-end learned CNNs.

Method	Top 1	Top 5	Params
AlexNet	56.9	80.1	61M
VGG-16 [12]	68.5	88.7	138M
Scat + Resnet-10 (ours)	68.7	88.6	12.8M
Resnet-18 (ours)	68.9	88.8	11.7M
Resnet-200 [41]	**78.3**	**94.2**	64.7M

Table 2. ILSVRC-2012 validation accuracy (single crop) of hybrid scattering and 10 layer resnet, a comparable 18 layer resnet, and other well known benchmarks. We obtain comparable performance using analogous amount of parameters while learning parameters at a spatial resolution of 28×28

Method	Accuracy
Unsupervised Representations	
Roto-Scat + SVM [27]	82.3
ExemplarCNN [11]	84.3
DCGAN [29]	82.8
Scat + FC (ours)	**84.7**
Supervised and Hybrid	
Scat + Resnet (ours)	93.1
Highway network [35]	92.4
All-CNN [34]	92.8
WRN 16 - 8 [41]	95.7
WRN 28 - 10 [41]	**96.0**

Table 3. Accuracy of scattering compared to similar architectures on CIFAR10. We set a new state-of-the-art in the unsupervised case and obtain competitive performance with hybrid CNNs in the supervised case.

4.1. Deep Hybrid CNNs on ILSVRC2012

We showed in the previous section that a SLE followed by FC layers can produce results comparable with the AlexNet [19] on the Imagenet classification task. Here we consider cascading the scattering transform with a modern CNN architecture, such as Resnet [41, 13]. We take the Resnet-18 [41], as a reference and construct a similar architecture with only 10 layers on top of the scattering network. We utilize a scattering transform with $J = 3$ such that the CNN is learned over a spatial dimension of 28×28 and a channel dimension of 651 (3 color channels of 217 each). The ResNet-18 typically has 4 residual stages of 2 blocks each which gradually decrease the spatial resolution [41]. Since we utilize the scattering as a first stage we remove two blocks from our model. The network is described in Table 4.

We use the same optimization and data augmentation procedure described in Section 3.1 but with learning rate drops at 30, 60, and 80. We find that, when both methods are trained with the same settings of optimization and data

Stage	Output size	Stage details
scattering	28×28	$J = 3, 651$ channels
conv1	28×28	[256]
conv2	28×28	$\begin{bmatrix} 256 \\ 256 \end{bmatrix} \times 2$
conv3	14×14	$\begin{bmatrix} 512 \\ 512 \end{bmatrix} \times 2$
avg-pool	1×1	$[14 \times 14]$

Table 4. Structure of Scattering and Resnet-10 used in imagenet experiments. Taking the convention of [41] we describe the convolution size and channels in the Stage details

Stage	Output size	Stage details
scattering	8×8, 24×24	$J = 2$
conv1	8×8, 24×24	16×k , 32×k
conv2	8×8, 24×24	$\begin{bmatrix} 32\times k \\ 32\times k \end{bmatrix} \times n$
conv3	8×8, 12×12	$\begin{bmatrix} 64\times k \\ 64\times k \end{bmatrix} \times n$
avg-pool	1×1	$[8 \times 8]$, $[12 \times 12]$

Table 5. Structure of Scattering and Wide ResNet hybrid used in small sample experiments. Network width is determined by factor k. For sizes and stage details if settings vary we list CIFAR-10 and then the STL-10 network information. All convolutions are of size 3×3 and the channel width is shown in brackets for both the network applied to STL-10 and CIFAR-10. For CIFAR-10 we use $n = 2$ and for the larger STL-10 we use $n = 4$.

augmentation, and when the number of parameters is similar (12.8M versus 11.7 M) the scattering network combined with a resnet can achieve analogous performance (11.4% Top 5 for our model versus 11.1 %), while utilizing fewer layers. The accuracy is reported in Table 2 and compared to other modern CNNs.

This demonstrates both that the scattering networks does not lose discriminative power and that it can be used to replace early layers of standard CNNs. We also note that learned convolutions occur over a drastically reduced spatial resolution without resorting to pre-trained early layers which can potentially lose discriminative information or become too task specific.

4.2. Hybrid Supervised and Unsupervised Representations on CIFAR-10

We now consider the popular CIFAR-10 dataset consisting of colored images composed of 5×10^4 images for training, and 1×10^4 images for testing divided into 10 classes. We perform two experiments, the first with a cascade of fully connected layers, that allows us to evaluate the scattering transform as an unsupervised representation. In a second experiment, we again use a hybrid CNN architecture with a ResNet built on top of the scattering transform.

For the scattering transform we used $J = 2$ which means the output of the scattering stage will be 8×8 spatially and 243 in the channel dimension. We follow the training procedure prescribed in [41] utilizing SGD with momentum of 0.9, batch size of 128, weigh decay of 5×10^{-4}, and modest data augmentation of the dataset by using random cropping and flipping. The initial learning rate is 0.1, and we reduce it by a factor of 5 at epochs 60, 120 and 160. The models are trained for 200 epochs in total. We used the same optimization and data augmentation pipeline for training and evaluation in both case. We utilize batch normalization techniques at all layers which lead to a better conditioning of the optimization [16]. Table 4.1 reports the accuracy in the unsupervised and supervised settings and compares them to other approaches.

In the unsupervised comparison we consider the task of classification using only unsupervised features. Combining the scattering transform with a NN classifier consisting of 3 hidden layers, with width 1.1×10^4, we show that one can obtain a new state of the art classification for the case of unsupervised features. This approach outperforms all methods utilizing learned and not learned unsupervised features further demonstrating the discriminative power of the scattering network representation.

In the case of the supervised task we compare to state-of-the-art approaches on CIFAR-10, all based on end-to-end learned CNNs. We use a similar hybrid architecture to the successful wide residual network (WRN) [41]. Specifically we modify the WRN of 16 layers which consists of 4 convolutional stages. Denoting the widening factor, k, after the scattering output we use a first stage of $32 \times k$. We add intermediate 1×1 to increase the effective depth, without increasing too much the number of parameters. Finally we apply a dropout of 0.2 as specified in [41]. Using a width of 32 we achieve an accuracy of 93.1%. This is superior to several benchmarks but performs worse than the original ResNet [13] and the wide resnet [41]. We note that training procedures for learning directly from images, including data augmentation and optimization settings, have been heavily optimized for networks trained directly on natural images, while we use them largely out of the box we do believe there are regularization techniques, normalization techniques, and data augmentation techniques which can be designed specifically for the scattering networks.

4.3. Limited samples setting

A major application of a hybrid representation is in the setting of limited data. Here the learning algorithm is limited in the variations it can observe or learn from the data, such that introducing a geometric prior can substantially improve performance. We evaluate our algorithm on the

Method	100	500	1000
WRN 16-8	34.7 $\pm$ 0.8	46.5 $\pm$1.4	60.0 $\pm$1.8
Scat + WRN 12-8	**38.9 $\pm$ 1.2**	**54.7$\pm$0.6**	**62.0$\pm$1.1**

Table 6. Mean accuracy of a hybrid scattering in a limited sample situation on CIFAR-10 dataset. We find that including a scattering network is significantly better in the smaller sample regime of 500 and 100 samples.

limited sample setting using a subset of CIFAR-10 and the STL-10 dataset.

4.3.1 CIFAR-10

We take subsets of decreasing size of the CIFAR dataset and train both baseline CNNs and counterparts that utilize the scattering as a first stage. We perform experiments using subsets of 1000, 500, and 100 samples, that are split uniformly amongst the 10 classes.

We use as a baseline the Wide ResNet [41] of depth 16 and width 8, which shows near state-of-the-art performance on the full CIFAR-10 task in the supervised setting. This network consists of 4 stages of progressively decreasing spatial resolution detailed in Table 1 of [41]. We construct a comparable hybrid architecture that removes a single stage and all strides, as the scattering already down-sampled the spatial resolution. This architecture is described in Table 5. Unlike the baseline, refereed from here-on as WRN 16-8, our architecture has 12 layers and equivalent width, while keeping the spatial resolution constant through all stages prior to the final average pooling.

We use the same training settings for our baseline, WRN 16-8, and our hybrid scattering and WRN-12. The settings are the same as those described for CIFAR-10 in the previous section with the only difference being that we apply a multiplier to the learning rate schedule and to the maximum number of epochs. The multiplier is set to 10,20,100 for the 1000,500, and 100 sample case respectively. For example the default schedule of 60,120,160 becomes 600,1200,1600 for the case of 1000 samples and a multiplier of 10. Finally in the case of 100 samples we use a batch size of 32 in lieu of 128.

Table 6 corresponds to the averaged accuracy over 5 different subsets, with the corresponding standard error. In this small sample setting, a hybrid network outperforms the purely CNN based baseline, particularly when the sample size is smaller. This is not surprising as we incorporate a geometric prior in the representation.

4.3.2 STL-10

The SLT-10 dataset consists of colored images of size 96×96, with only 5000 labeled images in the training set divided equally in 10 classes and 8000 images in the test set. The larger size of the images and the small number of available samples make this a challenging image classification task. The dataset also provides 100 thousand unlabeled images for unsupervised learning. We do not utilize these images in our experiments, yet we find we are able to outperform all methods which learn unsupervised representations using these unlabeled images, obtaining very competitive results on the STL-10 dataset.

Method	Accuracy
Supervised methods	
Scat + WRN 19-8	**76.0 $\pm$ 0.6**
CNN[37]	70.1 $\pm$ 0.6
Unsupervised methods	
Exemplar CNN [11]	75.4 $\pm$ 0.3
Stacked what-where AE [43]	74.33
Hierarchical Matching Pursuit (HMP) [3]	64.5$\pm$1
Convolutional K-means Network [8]	60.1$\pm$1

Table 7. Mean accuracy of a hybrid CNN on the STL-10 dataset. We find that our model is better in all cases even compared to those utilizing the large unsupervised part of the dataset.

We apply a hybrid convolutional architecture, similar to the one applied in the small sample CIFAR task, adapted to the size of 96×96. The architecture is described in Table 5 and is similar to that used in the CIFAR small sample task. We use the same data augmentation as with the CIFAR datasets. We apply SGD with learning rate 0.1 and learning rate decay of 0.2 applied at epochs 1500,2000,3000,4000. Training is run for 5000 epochs. We use at training and evaluation the standard 10 folds procedure which takes 1000 training images. The averaged result for 10 folds is reported in Table 7. Unlike other approaches we do not use the 4000 remaining training image to perform hyper-parameter tuning on each fold, as this is not representative of typical small sample situations, instead we train the same settings on each fold. The best reported result in the purely supervised case is a CNN [37, 11] whose hyper parameters have been automatically tuned using 4000 images for validation achieving 70.1% accuracy. The other competitive methods on this dataset utilize the unlabeled data to learn in an unsupervised manner before applying supervised methods. To compare with [14] we also train on the full training set of 5000 images obtaining an accuracy of 87.6% on the test set, which is substantially higher than 81.3% reported in [14] using unsupervised learning and the full unlabeled and labeled training set. The competing techniques add several hyper parameters and require an additional engineering process. Applying a hybrid network is on the other hand straightforward and is very competitive with all the existing approaches, without using any unsupervised learning.

In addition to showing hybrid networks perform well

in the small sample regime these results, along with our unsupervised CIFAR-10 results, suggest that completely unsupervised feature learning on natural image data, for downstream discriminative tasks, may still not outperform supervised learning methods and pre-defined representations. One possible explanation is that in the case of natural images, learning in an unsupervised way more complex variabilities than geometric ones (e.g the rototranslation group), might be very challenging or possibly ill-posed.

5. Conclusion

This work demonstrates a competitive approach for large scale visual recognition, based on scattering networks, in particular for ILSVRC2012. When compared with unsupervised representation on CIFAR-10 or small data regimes on CIFAR-10 and STL-10, we demonstrate state-of-the-art results. We build a supervised Shared Local Encoder that permits the scattering networks to surpass other local encoding methods on ILSVRC2012. This network of just 3 learned layers permits analysis on the operation performed.

Our work also suggests that pre-defined features are still of interest and can provide enlightenment on deep learning techniques and to allow them to be more interpretable. Combined with appropriate learning methods, they could permit having more theoretical guarantees that are necessary to engineer better deep models and stable representations.

Acknowledgments

The authors would like to thank Mathieu Andreux, Matthew Blaschko, Carmine Cella, Bogdan Cirstea, Michael Eickenberg, Stéphane Mallat for helpful discussions and support. The authors would also like to thank Rafael Marini and Nikos Paragios for use of computing resources. We would like to thank Florent Perronnin for providing important details of their work. This work is funded by the ERC grant InvariantClass 320959, via a grant for PhD Students of the Conseil régional d'Ile-de-France (RDM-IdF), Internal Funds KU Leuven, FP7-MC-CIG 334380, an Amazon Academic Research Award, and DIGITEO 2013-0788D - SOPRANO.

References

[1] S. Bernstein, J.-L. Bouchot, M. Reinhardt, and B. Heise. Generalized analytic signals in image processing: comparison, theory and applications. In *Quaternion and Clifford Fourier Transforms and Wavelets*, pages 221–246. Springer, 2013.

[2] L. Bo, X. Ren, and D. Fox. Multipath sparse coding using hierarchical matching pursuit. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 660–667, 2013.

[3] L. Bo, X. Ren, and D. Fox. Unsupervised feature learning for rgb-d based object recognition. In *Experimental Robotics*, pages 387–402. Springer, 2013.

[4] Y.-L. Boureau, N. Le Roux, F. Bach, J. Ponce, and Y. LeCun. Ask the locals: multi-way local pooling for image recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2651–2658. IEEE, 2011.

[5] J. Bruna and S. Mallat. Audio texture synthesis with scattering moments. *arXiv preprint arXiv:1311.0407*, 2013.

[6] J. Bruna and S. Mallat. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1872–1886, 2013.

[7] J. Bruna, A. Szlam, and Y. LeCun. Learning stable group invariant representations with convolutional networks. *arXiv preprint arXiv:1301.3537*, 2013.

[8] A. Coates and A. Y. Ng. Selecting receptive fields in deep networks. In *Advances in Neural Information Processing Systems*, pages 2528–2536, 2011.

[9] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[10] I. Dokmanić, J. Bruna, S. Mallat, and M. de Hoop. Inverse problems with invariant multiscale statistics. *arXiv preprint arXiv:1609.05502*, 2016.

[11] A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, and T. Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 766–774, 2014.

[12] S. Han, J. Pool, J. Tran, and W. Dally. Learning both weights and connections for efficient neural network. In *Advances in Neural Information Processing Systems*, pages 1135–1143, 2015.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[14] E. Hoffer, I. Hubara, and N. Ailon. Deep unsupervised learning through spatial contrasting. *arXiv preprint arXiv:1610.00243*, 2016.

[15] M. Huh, P. Agrawal, and A. A. Efros. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016.

[16] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[17] J.-H. Jacobsen, E. Oyallon, S. Mallat, and A. W. Smeulders. Multiscale hierarchical convolutional networks. *arXiv preprint arXiv:1703.01775*, 2017.

[18] J. J. Koenderink and A. J. Van Doorn. The structure of locally orderless images. *International Journal of Computer Vision*, 31(2-3):159–168, 1999.

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

[20] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural

scene categories. In *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, volume 2, pages 2169–2178. IEEE, 2006.

[21] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3361–3368. IEEE, 2011.

[22] Y. LeCun, K. Kavukcuoglu, C. Farabet, et al. Convolutional networks and applications in vision. In *ISCAS*, pages 253–256, 2010.

[23] D. G. Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999.

[24] S. Mallat. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012.

[25] S. Mallat. Understanding deep convolutional networks. *Phil. Trans. R. Soc. A*, 374(2065):20150203, 2016.

[26] E. Oyallon. Building a regular decision boundary with deep networks. *arXiv preprint arXiv:1703.01775*, 2017.

[27] E. Oyallon and S. Mallat. Deep roto-translation scattering for object classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2865–2873, 2015.

[28] F. Perronnin and D. Larlus. Fisher vectors meet neural networks: A hybrid classification architecture. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3743–3752, 2015.

[29] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[30] J. Sánchez and F. Perronnin. High-dimensional signature compression for large-scale image classification. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1665–1672. IEEE, 2011.

[31] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *International journal of computer vision*, 105(3):222–245, 2013.

[32] T. Serre and M. Riesenhuber. Realistic modeling of simple and complex cell tuning in the hmax model, and implications for invariant object recognition in cortex. Technical report, DTIC Document, 2004.

[33] L. Sifre and S. Mallat. Rotation, scaling and deformation invariant scattering for texture discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1233–1240, 2013.

[34] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[35] R. K. Srivastava, K. Greff, and J. Schmidhuber. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015.

[36] M. Sugiura. *Unitary representations and harmonic analysis: an introduction*, volume 44. Elsevier, 1990.

[37] K. Swersky, J. Snoek, and R. P. Adams. Multi-task bayesian optimization. In *Advances in neural information processing systems*, pages 2004–2012, 2013.

[38] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

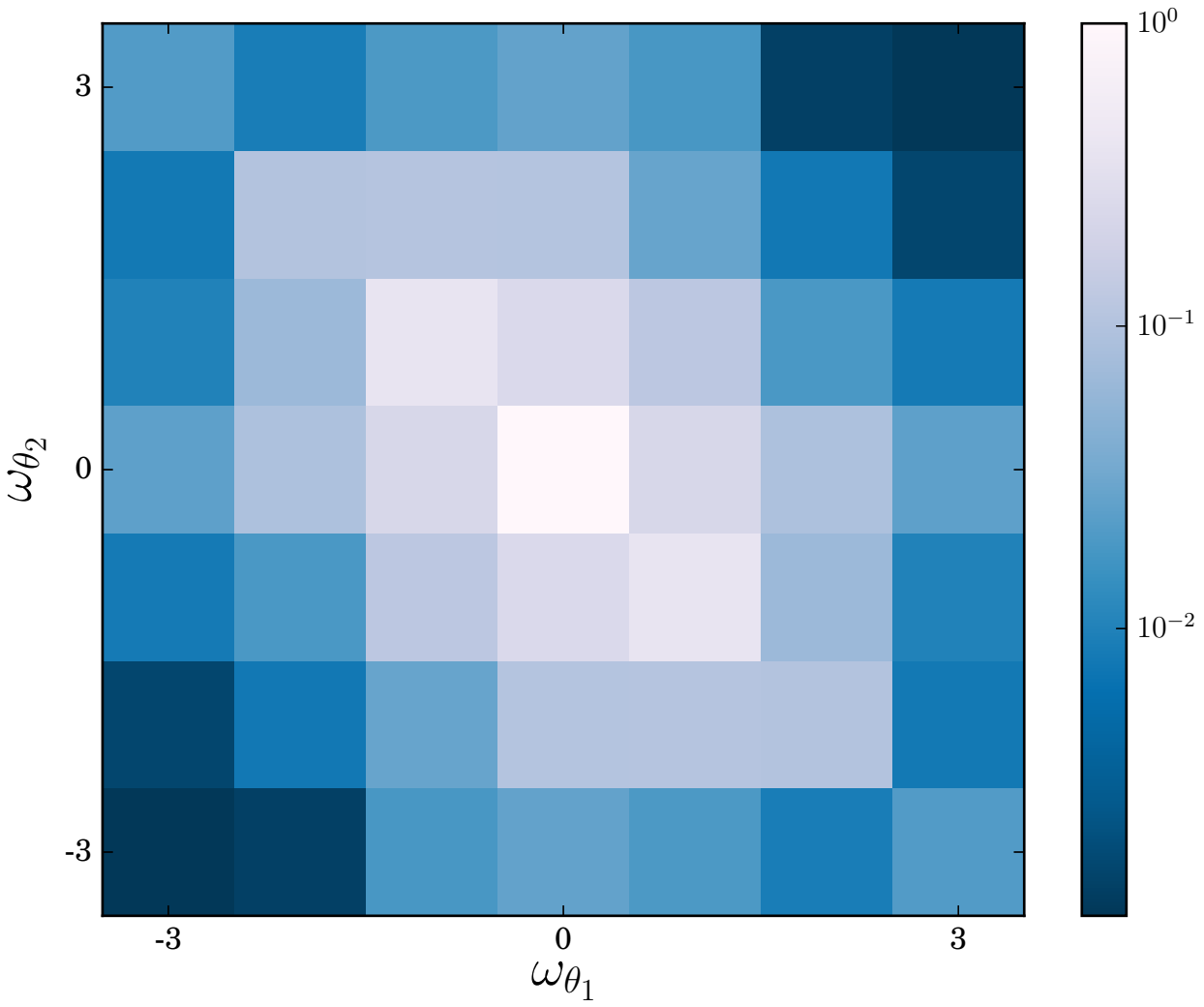
[39] I. Waldspurger. *These de doctorat de lEcole normale supérieure*. PhD thesis, École normale supérieure, 2015.

[40] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.

[41] S. Zagoruyko and N. Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

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

[43] J. Zhao, M. Mathieu, R. Goroshin, and Y. LeCun. Stacked what-where auto-encoders. *arXiv preprint arXiv:1506.02351*, 2016.


3
0
-3
ω_{θ_2}
-3
0
3
ω_{θ_1}
10^0
10^{-1}
10^{-2}