Generic Deep Networks with Wavelet Scattering

Edouard Oyallon
Ecole Normale Supérieure, 45 rue d’Ulm, 75005 Paris, France

Stéphane Mallat
Ecole Normale Supérieure, 45 rue d’Ulm, 75005 Paris, France

Laurent Sifre
Ecole Normale Supérieure, 45 rue d’Ulm, 75005 Paris, France

Abstract
We introduce a two-layer wavelet scattering network, which involves no learning, for object classification. This scattering transform computes a spatial wavelet transform on the first layer and a joint wavelet transform along spatial, angular and scale variables in the second layer. Image classification results are given on Caltech databases.

1. Introduction
Supervised training of deep convolution networks [LeCun et al., 1998] is clearly highly effective for image classification, as shown by results on ImageNet [Krizhevsky et al., 2012]. The first layers of the networks trained on ImageNet also perform very well to classify images in very different databases, which indicates that these layers capture generic image information [Zeiler & Fergus, 2013; Girshick et al., 2013; Donahue et al., 2013]. This paper shows that such generic properties can be captured by a scattering transform. Scattering transforms compute hierarchical invariants along groups of transformations by cascading wavelet convolutions and modulus non-linearities, along the group variables [Mallat, 2012]. Invariant scattering transforms to translations [Bruna & Mallat, 2013] and rotation-translations [Sifre & Mallat, 2012] have been applied to digit recognition and texture discrimination with important deformations. This paper defines a scattering transform computed over translation, rotation and scaling variables, implemented with a double layer convolution network. Image classification results are given on the Caltech-101 and Caltech-256 databases.

Submitted to the workshop track of International Conference on Learning Representations (ICLR), 2014

Figure 1. A scattering representation is computed by successively computing the modulus of wavelet coefficients with $|W_1|$, $|W_2|$, followed by an average pooling $\phi_J$.

2. Scattering along Translations, Rotations and Scales
A two-layer scattering transform is computed by cascading wavelet transforms and modulus non-linearities. The first wavelet transform $W_1$ filters the image $x$ with complex wavelets which are scaled and rotated. Taking the modulus of each complex coefficients defines the first scattering layer $U_1 x$. The modulus of a second wavelet transform $W_2$ computes the next layer $U_2 x$. Scattering coefficients $S_3 x$ are obtained with a final average pooling, as illustrated in Figure 1.

The first wavelet transform is defined from a mother wavelet $\psi^1(u)$, which is a complex function well localized in the image plane. This wavelet is scaled by $2^{-j_1}$, where $j_1$ is an integer or half-integer, and rotated by $\theta_1 = 2k\pi/K$ for $0 \leq k < K$:

$$\psi^1_{w_1}(u) = 2^{-2j_1} \psi^1(2^{-2j_1} r_{\theta_1} u)$$

where $w_1 = (\theta_1, j_1)$.

This wavelet transform filters an image $x(u)$, and we compute the modulus of the resulting wavelet coefficients:

$$U_1 x(u, w_1) = |x * \psi^1_{w_1}(u)| = \left| \sum_v x(u) \psi^1_{w_1}(u - v) \right| .$$

The spatial variable $u$ is subsampled by $2^{j_1-1}$. We write $u_1 = (u, w_1)$ the aggregated variable which indexes these first layer coefficients.
The next layer is computed with a second wavelet transform which convolves \( U_1 x(u_1) \) with separable wavelets along the spatial, rotation and scale variables \( u_1 = (u, \theta_1, j_1) \)

\[
\psi_{w_2}^2(u_1) = \psi_{j_2}^a(u) \psi_{k_2}^b(\theta_1) \psi_{l_2}(j_1).
\]

The index \( w_2 = (j_2, k_2, l_2) \) specifies the scales \( 2^{j_2}, 2^{k_2} \) and \( 2^{l_2} \) of these wavelets. We choose this wavelet family so that it defines a tight frame, and hence an invertible linear operator which preserves the norm. This wavelet family also includes the necessary averaging filters. The next layer of coefficients are defined for \( u_1 = (u, w_1) \) and \( w_2 = (k_2, l_2, m_2) \) by

\[
U_2 x(u_1, w_2) = |U_1 x \ast \psi_{w_2}^2(u_1)| = \left| \sum_{v_1} U_1 x(v_1) \psi_{w_2}^0(u_1 - v_1) \right|.
\]

According to most classification algorithms, invariance to illumination variability is implemented with a renormalization. The Euclidean norm of vectors of coefficients \( U_2 x(u_1, w_2) \) are set to 1 over spatial blocks of 4 by 4 coefficients.

A locally invariant translation scattering representation is obtained with a spatial pooling which averages the coefficients \( U_2 x \) over a window \( \phi_j(u) = 2^{-2J} \phi(2^{-J}u) \) along the spatial variable \( u \):

\[
S_3 x = U_2 x \ast \phi_J.
\]

This averaging is spatially subsampled at intervals \( 2^{J-1} \).

3. Numerical Classification Results

The classification performance of this double layer wavelet scattering representation is evaluated on the Caltech databases. All images are first renormalized to a fixed size of 150 by 150 pixels by a linear rescaling. The first wavelet transform \( W_1 \) is computed with Morlet wavelets (Bruna & Mallat 2013), over 2 octaves \( 1 \leq 2^{j_1} \leq 2^2 \) with \( K = 8 \) angles \( \theta_1 \). The second wavelet transform \( W_2 \) is computed with a two-dimensional Haar wavelet \( \psi^a \) over a range of spatial scales \( 2^{j_2} \leq 2^{k_2} \leq 2^5 \). We also use a Haar wavelet \( \psi^b \) over the angle variable, calculated over 3 octaves \( 1 \leq 2^{l_2} \leq 2^3 \). In this implementation, we did not use a wavelet along the scale variable \( j_1 \) because the first layer computes only over two scale values \( j_1 = 1, 2 \).

The final scattering coefficients \( S_3 x \) are computed with a spatial pooling at a scale \( 2^J = 64 \), as opposed to the maxima selection used in most convolution networks. These coefficients are renormalized by a standardization which subtracts their mean and sets their variance to 1. The mean and variance are computed on the training databases. Standardized scattering coefficients are then provided to a linear SVM classifier.

Almost state of the art classification results are obtained on Caltech-101 and Caltech-256, with a ConvNet (Zeiler & Fergus 2013) pretrained on ImageNet. Table 1 shows that with 7 layers it has an 85.5% accuracy on Caltech-101 and 72.6% accuracy on Caltech-256, using respectively 30 and 60 training images. The classification is performed with a linear SVM. In this work, we concentrate on the first two layers. With only two layers, the ConvNet performances drop to 66.2% on Caltech-101 and 39.6% on Caltech-256, and progressively increase as the number of layers increases. A scattering transform has similar performances as a ConvNet when restricted to 1 and 2 layers, as shown by Table 1. It indicates that major sources of classification improvements over these first two layers can be obtained with wavelet convolutions over spatial variables on the first layer, and joint spatial and rotation variables on the second layer. More improvements can potentially be obtained by adjusting the wavelet filtering along scale variables.

4. Conclusion

We showed that a two layer scattering convolution network, which involves no learning, provides similar accuracy on the Caltech databases, as double layers neural network pretrained on ImageNet. This scattering transform linearizes the variability relatively to translations and rotations and provides invariants to translations with an average pooling.

A scattering transform can also linearize scaling image variability but this is not used in these numerical evaluations. No optimization was performed on the choice of wavelets. Many improvements can be brought to these first scattering layers. However, these preliminary experiments indicate that wavelet scattering transforms may allow to un-
understand the behavior of the first two convolution network layers for image classification.

References


