Deep Roto-Translation Scattering for Object Classification

Edouard Oyallon[1], Stéphane Mallat[1]
[1]Département Informatique, Ecole Normale Supérieure

A scattering transform is a deep convolutional network computed with predefined complex wavelet filters. It has been sucessfully applied to texture classification [5] and digit recognition[1]. In this paper, we introduce a separable scattering along translation and rotation group, for object classification in images. Applying a supervised SVM classifier with a feature selection gives an accuracy similar to state-of-the-art unsupervised learning algorithms, on Caltech and CIFAR image datasets.

Image classification requires to find sets of features which are discriminative and do not suffer from the high variance resulting from the high image dimensionality. We show that building stable invariants adapted to translation, rotation and image deformations leads to feature sets which perform as well as unsupervised learning.

Scattering coefficients are computed with a cascade of wavelet transforms, and modulus non-linearities. Wavelets separate the image information along multiple scales and angles which provides a representation which is stable to deformations. Translation invariance at a scale 2^J is obtained with an average pooling with a dilated Gaussian $\phi_J(p) = 2^{-2J}\phi(2^{-J}p)$.

A scattering network computes the modulus of wavelet coefficients at multiple scales. They are calculated with a complex $2D$ Morlet (nearly Gabor) wavelet ψ, which is dilated by 2^j and rotated by 8 angles θ:

$$\psi_{j,\theta}(p) = 2^{-2j}\psi(2^{-j}r_{-\theta}\,p)\ .$$

A wavelet convolution at a scale 2^j can be implemented with a cascade of j filtering and subsampling, so the scale index j corresponds to the network depth. First order internal wavelet coefficients are calculated at all scales $0 \leq j \leq J$:

$$x^1_j(p,\theta) = |x \star \psi_{j,\theta}(p)|.$$

A separable roto-translation scattering computes second order internal coefficients by filtering each $x^1_j(p,\theta)$ along the spatial variable p and the angular variale θ, with a $3D$ separable complex wavelet:

$$\psi_{j,\beta,k}(p,\theta) = \psi_{j,\beta}(p)\,\overline{\psi}_k(\theta)\ .$$

$\overline{\psi}_k(\theta) = 2^{-k}\overline{\psi}(2^{-k}\theta)$ is one-dimensional angular wavelet. Second order internal wavelet coefficients recombine spatial and angular information with a 3D convolution of $x^1_{j_1}(p,\theta)$ along (p,θ), for any $0 \leq j_1 < j$:

$$x^2_j = |x^1_{j_1} \star \psi_{j,\beta,k}|.$$

The angular filtering provides a sensitivity to angular variations which improves classification.

The last layer J of this scattering network outputs a spatial average pooling at the scale 2^J, of all internal coefficients, subsampled at spatial intervals 2^J:

$$\left\{x \star \phi_J\,,\, x^1_j \star \phi_J\,,\, x^2_j \star \phi_J\right\}_{1 \leq j \leq J}\ .$$

It includes a spatial averaging of the input image image x, of first order wavelet modulus images x^1_j, and of second order coefficients x^2_j, at all scales $2^j \leq 2^J$.

For an image of $P = 256^2$ coefficients, at a scale $2^J = 2^6$, there is about 50 averaged coefficient $x \star \phi_J$. For a wavelet transform computed over 8 angles θ, there is nearly $2\,10^3$ coefficients $x^1_j \star \phi_J$, and about $90\,10^3$ coefficients $x^2_j \star \phi_J$. It is a rich set of locally invariant descriptors, providing joint information across scales and angles.

We introduce a supervised classifier, which first performs a feature selection with an orthogonal least square. Scattering coefficients are highly correlated. The orthogonal least square is a supervised greedy algorithm, which selects $M \approx 3\,10^3$ features with labeled training data. It decorrelates the selected scattering features with a Gram-Schmidt orthogonalization. This selection can be interpreted as a fully connected layer with M nodes at the output. The final classification is performend with a Gaussian kernel SVM on these M coefficients.

Dataset	ScatNet	Unsupervised	Supervised
Caltech-101	79.9	82.5	91.4
Caltech-256	43.6	50.7	70.6
CIFAR-10	82.3	83.1	91.8
CIFAR-100	56.8	60.8	65.4

Table 1: Classification accuracy of a roto-translation scattering compared to state of the art unsupervised and supervised feature learning.

The resulting classification accuracy is evaluated on CIFAR and CalTech data bases. A scattering representation is a set of predefined features calculated with wavelet filters. It yields much better accuracy results than any existing algorithm using predefined features, such as SIFT type features or random filters in deep networks, which are not optimized by learning. Table 1 compares this classification accuracy with unsupervised feature learning, without data augmentation, and with supervised feature learning algorithms. The best results are obtained with supervised feature learning, using deep networks which are trained on ImageNet or by using data augmentation technics.

Table 1 shows that predefined scattering features gives a classification accuracy of the same order as state of the art unsupervised feature learning. The same scattering classifier is used for the four data bases whereas results from different unsupervised learning algorithms relatively to each data basis are given in the table 1. No single unsupervised learning algorithm performs better than a scattering on all data bases. Supervised deepnetwork with data augmentation lead to the state-of-the-arts results on those four datasets [3].

A scattering transform is computed with convolutions along groups of transformations, which are responsible for important image variabilities. This paper concentrates on translations and rotations, but it can similarly be extended to any other group. Improving results requires to consider other source of variabilities and invariants, for example across color channels or across scales, which are not recombined in this architecture. The hard part is to identify the important group of variability for improving classification. It seems that supervised deep network classifiers are able to identify them.

[1] Joan Bruna and Stéphane Mallat. Invariant Scattering Convolution Networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1872–1886, 2013.

[2] Sheng Chen, Colin FN Cowan, and Peter M Grant. Orthogonal least squares learning algorithm for radial basis function networks. *Neural Networks, IEEE Transactions on*, 2(2):302–309, 1991.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *NIPS*, pages 1106–1114, 2012.

[4] Stéphane Mallat. Group Invariant Scattering. *CoRR*, abs/1101.2286, 2011.

[5] Laurent Sifre and Stéphane Mallat. Rotation, Scaling and Deformation Invariant Scattering for Texture Discrimination. In *CVPR*, pages 1233–1240. IEEE, 2013.

[6] Matthew D. Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. *CoRR*, abs/1311.2901, 2013.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.