Scattering Convolutional Networks for Image Classification and Analysis

Quentin Le Roux

Université Côte d'Azur, France

Abstract. A scattering network is a cascade of wavelet transforms with non-linearities. They produce invariants: feature maps invariant to isometries (rotation, translation) and small deformations. Comparable to convolutional neural networks, scattering networks display mathematical certainty and explanability, and have been shown to surpass learned architectures in some areas such as subsampled dataset classification. This report covers the broad concept of scattering networks and introduces their most recent developments.

Keywords: Signal Representation · Wavelet · Wavelet Transforms · Scattering Networks · Convolutional Neural Networks

1 Signal Decomposition

1.1 Origins: the Fourier Transform

The Fourier Transform (FT) is a signal decomposition method that extracts frequency and amplitude information from stationary signals while obfuscating frequency-time information.

$$\hat{f}(\omega) = \int_{-\infty}^{+\infty} f(t) e^{-2\pi i \omega t} dt \tag{1}$$

t, time; ω, a frequency; $f(t)$, a signal intensity vs. time function

The Short-Time Fourier Transform (STFT) helps deal with non-stationary signals[23]. It extracts localized frequency information in a given time window, relying on the assumption that a non-stationary signal still presents stationary subparts. As such, the FTFT computes the FT over fixed slices of a given signal.

$$\hat{f}(\omega, \tau) = \int_{-\infty}^{+\infty} f(t) w(t - \tau) e^{-2\pi i \omega t} dt \tag{2}$$

w, a window function; τ, a translation parameter

However, fixed-length time windows result in interfering frequency and time resolutions. Short windows are best suited for high frequencies (effective time but poor frequency resolutions) while wide windows are best suited for low frequencies (effective frequency but poor time resolutions).

1.2 Wavelets, the Wavelet Transform, and multilevel decomposition

Wavelets are small, localized wave-like functions[23] parametrized by two factors: window size and scaling (a large scaling yields better frequency resolution and a short one better time resolution). Based on them, the Wavelet Transform (WT) improves on the STFT's time-frequency trade-off.

$$\hat{f}(s,\tau) = \frac{1}{\sqrt{|s|}} \int_{-\infty}^{+\infty} f(t)\psi(\frac{t-\tau}{s})dt \tag{3}$$

ψ, a wavelet function; s, the scale; τ, the window size

Computing wavelet coefficients $\hat{f}(s,\tau)$ at each scale and window is inefficient, however. This is solved with the Discrete Wavelet Transform (DWT):

$$\hat{f}(s,\tau) = \frac{1}{\sqrt{|s|}} \sum_{m=0}^{p-1} f(t_m)\psi(\frac{t_m-\tau}{s}) \tag{4}$$

$s = 2^{-j}$ the dyadic dilation; $\tau = k2^{-j}$ the dyadic position

j, the scale index; k, the wavelet transform signal index

The DWT ends up providing an accurate and finer signal windowing and time-frequency trade-off compared to other decomposition methods (See Fig. 1).

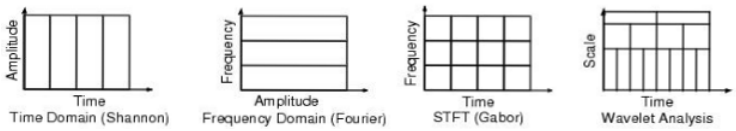

Fig. 1. DWT allows efficient retrieval of low and high-frequency information[31].

2 From the Wavelet Transform to Scattering Networks

2.1 The importance of invariance in signal analysis

Achieving invariance is about reducing intra-class variance within a dataset. Applied to signals, it implies constructing feature maps invariant to translation and transposition (audio), or rotation and scaling (images) for instance.

$$\forall c \in R,\ x_c(t) = x(t-c) \qquad \text{(translation)} \tag{5}$$

$$\forall \tau, x_\tau(t) = x(t-\tau(t)) \qquad \text{(deformation)} \tag{6}$$

Invariance can further be described as a Lipschitz continuity condition[3][22]:

$$||\Phi(x_\tau) - \Phi(x)|| \leq C \sup_t |\nabla(t)|.||x|| \quad (\sup_t |\nabla\tau(t)|, \text{ the deformation size}) \tag{7}$$

Invariance to isometries and deformation is key in signal analysis as signals are characterized by non-rigid deformations – The MNIST dataset provides such an example[17]. As such, dealing with high intra-class variance is a necessary step to compute and learn effective signal representations [2][13][30].

It has been shown that the STFT is only robust to small deformations[3]. It erases high frequencies and thus is not suitable to build robust invariants. With its localized characteristic, the WT can achieve invariance, albeit imperfectly, to both translation and deformation when compared to the STFT[3][22].

2.2 The Wavelet Transform and invariance

The WT solves the high frequency issue of previous feature mapping methods such as the Scale-Invariant Feature Transform (SIFT)[3] while achieving the same purpose: to build lower-dimensional manifolds to be used as inputs to supervised or unsupervised models (e.g. classification with Support Vector Machines (SVM) or clustering with Gaussian Mixture Models (GMM)[26][27]).

To achieve invariance, the WT is duplicated (via rotation and dilation of the underlying mother wavelet $\psi(t)$) to cover the entire frequency domain of a signal x. This operation yields $W_x(t)$, the set of convolutions $x \star \psi_\lambda(t)$ of the signal such that the lowest frequencies are recorded via a low-pass filter $x \star \phi$[22]:

$$W_x(t) = \left\{x \star \phi, x \star \psi_\lambda(t)\right\}_\lambda \quad (8)$$
$$\text{with } \lambda, \text{ a rotation tuning parameter}$$

Such sets are not translation-invariant per se due to an averaging operation at high frequencies that induces a loss of information in order to build invariants.

2.3 Scattering Transforms

The loss of information induced by computing $W_x(t)$ via the WT is addressed by introducing the non-linear modulus operator and an iterative process that relies on a function U_x called the scattering propagator (SP)[3][22] of a signal x:

$$U_x(t) = \left\{x \star \phi, |x \star \psi_\lambda(t)|\right\}_\lambda \quad (9)$$
$$\forall z \in C,\ z = a + ib;\ |z| = \sqrt{a^2 + b^2}$$

Scattering Transforms (ST) are an accumulated composition of SP U_x. Each SP outputs the low-pass filter $x \star \phi$ as an intermediary output, and passes the result of its modulus operation $|x \star \psi_\lambda(t)|$ onto the next composed SP, resulting in p paths constructed over the set of rotations λ of $\psi(t)$ (See Fig. 3):

$$\forall \text{ path } p = (\lambda_1, ..., \lambda_m) \text{ of order } m$$
$$S[p]x(t) = |\ldots||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|\ldots| \star \psi_{\lambda_m}| \star \phi(t) \quad (10)$$

This results into a ST tree S of order m parametrized by a scale J and L different orientations for the underlying mother wavelet function $\psi(t)$[24]. A ST

of order 1 is similar to the SIFT[18]. Furthermore, it has been shown that the paths of a ST tree S can be intelligently pruned to accelerate computation. As high-energy features are concentrated along specific paths p, a reduced ST tree S can compute the invariants of an n-sized signal x in $\mathcal{O}(n\ log\ n)$ time[3].

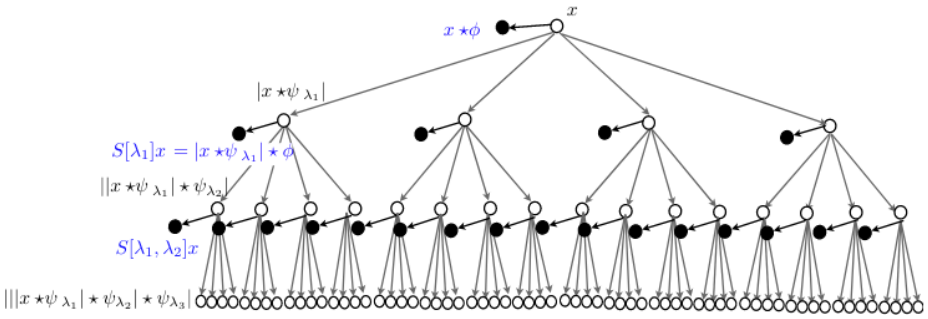

Fig. 2. Graphical representation of a ST tree of order 3[22].

Computing the ST tree S (a simultaneous feature propagation and low-pass filtering[29]) results in deformation, rotation and translation-invariant feature mappings[20][21][22]. Such representations happen to be sparse, accurate, and stable representations with relatively few parameters to characterize them[3].

2.4 Scattering Networks

ST were pioneered by Joan Bruna and Stéphane Mallat in 2012, shortly after convolutional neural network architectures (CNN) rose to prominence around 2010[16]. From their inception, ST have also been described as Scattering Convolutional Networks (SN): a non-learned wavelet-cascade comparable to their learned CNN counterparts[1][3][21][22][29].

The parameters of a SN are fixed and tightly tied to the shape of the input signals to transform and the desired output shape[1][20][21][22] (See Table 1).

Table 1. Parameters for 1- and 2-dimensional SN

Parameter	SN dimension	Explanation
Q	1D	The number of wavelets per octave
J	1D, 2D	The number of scales to cover the input signal
L	2D	The number of angles to cover the input signal

[1] "The scattering transform is defined as a complex-valued convolutional neural network whose filters are fixed [...] and the non-linearity is a complex modulus. Each layer is a wavelet transform, which separates the scales of the incoming signal. [...] The result is a reduction of variance and a stability to additive noise[1]."

Given a N-sized dataset, the input of a 1-dimensional SN is of shape (N, T) with T the length of the input signal. Its output is of shape $(N, P, \frac{T}{2^J})$, with $P \propto 1 + JQ + J(J-1)\frac{Q}{2}$ the number of scattering coefficients.

The input of a 2-dimensional SN is of shape (N, C, W, H) with C the number of channels, and W and H the width and height of the signal respectively. Its output is of shape $(N, C, 1 + LJ + \frac{L^2 J(J-1)}{2}, \frac{W}{2^J}, \frac{H}{2^J})$[3]. Provided as a textbook example, the output of a SN of order 2 on an element from the MNIST dataset is reproduced in Fig. 3.

Paired with a SVM or a PCA, SN were found to outperform CNN on texture discrimination tasks, only necessitating to be of order 2[21][22] (a higher order m only yields marginal error improvements). SN were found to also be more robust in specific cases such as subsampled dataset classification (see Table 2).

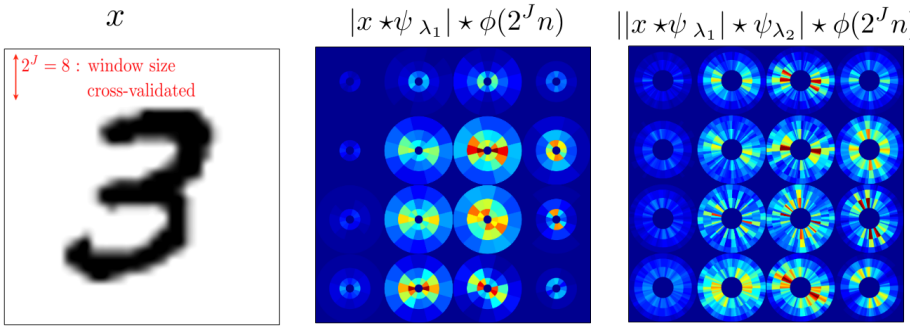

Fig. 3. SN output ($m = 2$, $J = 3$, and $L = 12$) of a MNIST dataset entry [19][22].

Most state-of-the-art implementations before 2018 were applied to texture discrimination databases[10][15][33] and the MNIST. It was noted early on that more complex datasets like Caltech101[9] may display complex variances where learned approaches could be better suited for the time being[3].

Table 2. Classification error of a SN against a then state-of-the-art CNN[16] on the MNIST dataset, reproduced from [3] (2013).

Training Size	PCA	SVM	SN with PCA ($m = 2$)	SN with SVM ($m = 2$)	CNN
300	14.5	15.4	**4.7**	5.6	7.18
1000	7.2	8.2	**2.3**	2.6	3.21
2000	5.8	6.5	**1.3**	1.4	2.53
5000	4.9	4	**1.03**	1.4	1.52
10000	4.55	3.11	0.88	1	**0.85**
20000	4.25	2.2	0.79	**0.58**	0.76
40000	4.1	1.7	0.74	**0.53**	0.65
50000	4.3	1.4	0.7	**0.43**	0.53

3 State-of-the-Art Applications of Scattering Networks

3.1 Hybrid Networks

Given that the first layers in a learned CNN display filters with properties similar to the WT[22], hybrid networks (HN) were first proposed in 2012[3] then fully implemented in 2018[24].

The main advantage of a HN (where the first layers of a neural network are replaced by a SN) is the production of fixed local encodings that reduce the number of learned parameters in a network, allowing for faster and lighter learning with only marginal increases in error rates. HN (coupled with either fully-connected layers or CNN) were found to rival learned methods[24] in classification and unsupervised tasks on the STL-10[6], ImageNet[7], and CIFAR-10[14] datasets. The second advantage is the tried robustness to data subsampling[24].

3.2 Max-Pooling layer variant

An extension to HN, scattering-maxp networks were recently introduced[29]. They rely on a continuous max-pooling function as the non-linearity at the SP level to further reduce the number of learnable network parameters.

The single paper[29] demonstrated an up to 8-fold reduction in the parameter size of HN architectures compared to fully-learned variants (up to 12 folds when compared to VGG-16 [28]). Even so, only small performance reduction were observed on the Caltech101 and Caltech256[12] datasets.

3.3 Image reconstruction and generation

It has been supposed[4][5][8] and recently shown with HN architecture[24] that SN feature mappings can be used in image reconstruction and generation tasks. A Scattering Deep Convolutional Generative Adversarial Network in a dataset's scattering space was demonstrated on CIFAR-10 albeit with a lower performance than with fully-learned methods (likely tied to non-surjectivity issues[24]).

3.4 Parametric Scattering Networks

Backpropagation through a SN has recently been demonstrated with Morlet wavelets (The derivations are not reproduced here but can be found in [11]):

$$\psi_{\sigma,\theta,\xi,\gamma}(u) = \exp\big(i\xi(u_1 cos(\theta) + u_2 sin(\theta)) - \beta\big).\exp\big(-\frac{\varphi(u)}{2\sigma^2}\big) \tag{11}$$

$$\varphi(u) = ||\begin{pmatrix} cos(\theta) & sin(\theta) \\ \gamma sin(\theta) & -\gamma cos(\theta) \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}||^2 \tag{12}$$

β, a normalization constant, ξ, the frequency scale; γ, the slant

σ, the Gaussian window scale; θ, the global orientation

This allows learning the parameters of each WT within a SN (4 parameters per WT in the case of Morlet wavelets) or a HN thanks to the chain rule. The

paper[11] demonstrated improved performance on all HN architectures (with either linear or residual downstream layers), and especially with subsampled datasets (tested with CIFAR-10, KTH-TIPS, and COVIDx CRX-2[32]).

3.5 Possible areas of research

No known theoretical guarantee currently exists that the error of an image reconstruction process based on scattering coefficients (as part of an HN) converges [24] – gradient descent has so far been performed ad-hoc with auto-differentiation tools such as PyTorch[25].

Meanwhile, exploring the interpretability of shared local encoders (a cascade of pointwise convolutions with a SN as input layer) could be of help to better understand the early layers of CNN [24].

4 Conclusion

SN are a small but fertile ground for innovation in the field of neural networks and feature learning. Providing robust applications and strong mathematical certainties to better understand CNN, their recent implementations show promises that could orient future research in the wider field of Deep Learning.

References

1. Andreux M., Angles T., Exarchakis G., Leonarduzzi R., Rochette G., Thiry L., Zarka J., Mallat S., Andén J., Belilovsky E., Bruna J., Lostanlen V., Hirn M. J., Oyallon E., Zhang S., Cella C., Eickenberg M.: Kymatio: scattering transforms in Python. arXiv preprint:1812.11214. (2019).
2. Bajcsy, R., Kovacic, S.: Multi-resolution elastic matching. In: Computer vision graphics and image processing, vol. 46, Issue 1. (1989)
3. Bruna, J., Mallat, S.: Invariant scattering convolution networks. In: IEEE transactions on pattern analysis and machine intelligence, 35(8):1872-1886. (2013)
4. Bruna, J.: Scattering representations of recognition. Polytechnique X. (2013)
5. Bruna, J., Mallat, S.: Audio texture synthesis with scattering moments. arXiv preprint:1311.0407. (2013)
6. Coates, A., Lee, H., Ng A.Y.: An analysis of single layer networks in unsupervised feature learning. In: AISTATS. (2011)
7. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: IEEE CVPR. (2009)
8. Dokmanić, I., Bruna, J., Mallat, S., de Hoop, M.: Inverse problems with invariant multiscale statistics. arXiv preprint:1609.05502. (2016)
9. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In: IEEE CVPR, workshop on generative-model based vision. (2004)
10. Fritz, M., Hayman, E., Caputo, B., Eklundh, J.O.: The KTH-TIPS database. In: CVAP dept. of Numerical Analysis and Computer Science. (1999)
11. Gauthier, S., Thérien, B., Alsène-Racicot, L., Rish, I., Belilovsky, E., Eickenberg, M., Wolf, G.: Parametric scattering networks. arXiv preprint :2107.09539. (2021)

12. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset. California Institute of Technology. (2007)
13. Keysers, D., Deselaers, T., Gollan, C., Ney, H.: Deformation models for image recognition. In: IEEE trans. of PAMI. (2007)
14. Krizhevsky, A.: Learning multiple layers of features from tiny images. University of Toronto. (2009)
15. Lazebnik, S., Schmid, C., Ponce, J.: A sparse texture representation using local affine regions. In: IEEE trans. of PAMI, vol. 27, no. 8, pp. 1265-1278. (2005)
16. LeCun, Y., Kavukcuoglu, K., and Farabet, C.: Convolutional networks and applications in vision. In: Circuits and Systems (ISCAS), IEEE, pp. 253-256. (2010)
17. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. In: proc. of the IEEE, 86(11):2278-2324. (1998)
18. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. In: International Journal of Computer Vision, 60, pp. 91-101. (2004)
19. Maison, J.: Image classification with scattering networks and convolutional networks. github.com/Jonas1312/scattering-networks. Last accessed 27 Nov 2021
20. Mallat, S.: Group invariant scattering. In: Communications on Pure and Applied Mathematics, 65(10):1331-1398. (2012)
21. Mallat, S., Sifre, L.: Rotation, scaling and deformation invariant scattering for texture discrimination. In: IEEE CVPR, pp. 1233-1240. (2013)
22. Mallat, S.: Scattering invariant deep networks for classification. youtu.be/4eyUReyIPXg, youtu.be/Gb8uaQn12Gk. Lecture at UCLA, Institute for Pure and Applied Mathematics. (2012). Last accessed 25 Nov 2021
23. Nicoll, A.: The wavelet transform for beginners, youtu.be/kuuUaqAjeoA. (2020). Last accessed 22 Nov 2021
24. Oyallon, E., Zagoruyko, S., Huang, G., Komodakis, N., Lacoste-Julien, S., Blaschko, M., Belilovsky, E.: Scattering networks for hybrid representation learning. In: IEEE transactions on pattern analysis and machine intelligence, 41(9) :2208–2221. (2018)
25. Paszke, A., Gross, S., Chintala, S., Chanan, G. Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in PyTorch. NIPS Workshop. (2017)
26. Sánchez, J., Perronnin, F.: High-dimensional signature compression for large-scale image classification. In: IEEE CVPAR, pp. 1665-1672. (2011)
27. Sánchez, J., Perronnin, F., Mensink, T., Verbeek, J.: Image classification with the Fisher vector: theory and practice. In: International Journal of Computer Vision, 105(3):222-245. (2013)
28. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint:1409.1556. (2014)
29. Taekyung, K., Youngmi, H.: Deep scattering network with max-pooling. In: IEEE Signal Processing Society SigPort. (2021)
30. Trouve, A., Younes, L.: Local geometry of deformable templates. In: SIAM Journal on Mathematical Analysis, vol. 37, Issue 1. (2005)
31. Ukil, A., Zivanovic, R.: Abrupt change detection in power system fault analysis using adaptive whitening filter and wavelet transform. In: Electric Power Systems Research, vol. 76, issues 9–10, pp. 815-823. (2006)
32. Wang, L., Wong, A.: COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. arXiv preprint:2003.09871. (2020)
33. Xu, Y., Ji, H., Fermüller: A projective invariant for texture. In: IEEE CVPR, pp. 1932-1939. (2006)