Scattering representation

Geert Kapteijns

November 27, 2017

Abstract

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

1 Introduction

Let me start off by saying this is an informal document meant to make my research efforts of the last three months accessible. I have made no huge effort to cite the first paper to establish a concept or even to cite well-established results at all.

The challenge at hand is, in the broadest sense, to use statistical models and computational power to improve outcome for ischemic stroke patients. This note focuses on using image data to aid radiologists in the process of deciding, in the earliest possible stage, whether to use intra-arterial therapy (a relatively new endovascular, catheter-based treatment.)

I have limited my study to finding image features in 3D NCCT scans that are relevant for this decision. I believe that, despite popularity and exceptional engineering results in many areas (easily outperforming "hand-crafted" algorithms), deep learning is not the best tool for this problem (and I suspect many other problems in the medical domain), because medical images are high-dimensional, training data is scarce and there is not always a clearly defined ground truth.

What I have tried here is to use a different approach, namely one where no model is trained for feature extraction, but which uses an image representation called the scattering transform, that is known to have certain desirable properties, e.g. it preserves high-frequency information, is translation invariant over a tunable window and stable to deformations. This image representation may, in association with a simple classifier like an SVM, be used in classification problems.

In the classification experiments, I have chosen to focus on classifying brain hemispheres as being unaffected or affected by stroke, using the MRCLEAN registry. While this is not a clinically relevant problem (classifying which hemispheres is affected is a trivial task in most cases?) it serves as a test for the scattering representation, as at least there is an unequivocal ground truth. Though not done in this work, the learned features may be visualised to aid radiologists in giving an ASPECTS score or otherwise diagnosing the seriousness of the stroke. It is also possible, if a good delineation of the ASPECTS regions is available, to directly train a classifier on the scattering representation of these regions (though expert's opinion on whether a region is affected or not is not unequivocal, making it difficult to attain super-expert performance).

2 Scattering transform

I will in the briefest possible way state what the scattering transform is. For details refer to

[**anden2014deep**, **bruna2013invariant**].

A wavelet transform is defined by convolving a d-dimensional signal $y(\boldsymbol{x})$ with scaled and rotated versions of a mother wavelet $\psi_{a^j,r}(\boldsymbol{x})$, with $j \in \mathbb{Z}$ and rotations $r \in SO(d)$ (rotation group in d dimensions). $d = 3$ in our case, since we are dealing with 3D images. $a = 2$ is common for image analysis. We will describe in the next section which mother wavelet we use in practice, and how to correctly choose a finite number of length scales and rotations r (which is not trivial in 3D).

A wavelet of dilation a^j and orientation r looks like

$$\psi_{a^j,r}(\boldsymbol{x}) = a^{-dj}\psi(a^{-j}r\boldsymbol{x}) \tag{1}$$

where the normalisation a^{-dj} is chosen such that the energy of the mother wavelet is conserved

$$\int_{\mathbb{R}^d} d\boldsymbol{x}|\psi_{a^j,r}(\boldsymbol{x})| = \int_{\mathbb{R}^d} d\boldsymbol{x}|\psi(\boldsymbol{x})|. \tag{2}$$

Translationally invariant coefficients (called scattering coefficients) of $y(\boldsymbol{x})$ that are stable to small deformations are obtained by taking the modulus and taking a spatial average:

$$\left\|y \star \psi_{a^j,r}\right\|_1 = \int d\boldsymbol{x}|y \star \psi_{a^j,r}|. \tag{3}$$

The signals $|y\star\psi_{a^j,r}|$ are themselves unstable, and in averaging (or equivalently, removing all non-zero frequencies) of $|y\star\psi_{a^j,r}|$ information is lost. To remedy this, we can perform a second wavelet transform on all first-order transforms, yielding second-order scattering coefficients

$$\left\||y \star \psi_{a^{j_1},r_1}| \star \psi_{a^{j_2},r_2}\right\| = \int d\boldsymbol{x}||y \star \psi_{a^{j_1},r_1}| \star \psi_{a^{j_2},r_2}| \tag{4}$$

for all j_1, j_2, r_1, r_2. One can keep iterating this transform to recover more lost information in the form of higher-order coefficients, but in practice two layers is sufficient for most tasks (luckily so, because the computational resources required scales exponentially in the number of layers).

To make notation easier, we define $\lambda = (j, r)$ and

$$U[\lambda_1, \dots \lambda_m]y = |||y \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \dots \psi_{\lambda_m}| \tag{5}$$

i.e. an unaveraged signal at the mth layer. The scattering coefficients at the mth layer are then written

$$\left\{\bar{S}y(\lambda_1, \dots \lambda_m) = \int d\boldsymbol{x}U[p]y\right\}_{\lambda_i \in \Lambda} \tag{6}$$

In practice it is often better not to average over the entire signals $U[p]$ (where we write $p = \lambda_1, \dots \lambda_m$ a *path*), but compute coefficients that are approximately translation invariant over lengths a^J. This is achieved by computing the wavelet transforms only at scales $j \leq J$ and averaging with a low-pass (blurring) filter (in practice a Gaussian) of support a^J, denoted by ϕ_{a^J}

$$S[p]y(\boldsymbol{x}) = (U[p] \star \phi_{a^J})(\boldsymbol{x}). \tag{7}$$

$\{S[\lambda_1, \dots, \lambda_m]\}_{\lambda_i \in \Lambda}$ are called the windowed scattering coefficients at layer m. The correct maximum length scale a^J should be chosen based on knowledge about the input data or by cross-validation.

3 Implementation details

What follows are some considerations for choosing the mother wavelet and rotations r.

3.1 Construction of the mother wavelet

For the scattering representation to be stable to additive noise and contain all high frequency information, it must satisfy the Littlewood-Paley condition

$$(1 - \epsilon) \leq A(\boldsymbol{\omega}) \leq 1 \qquad \forall \boldsymbol{\omega} \in \mathbb{R}^d \tag{8}$$

with

$$A(\boldsymbol{\omega}) = \left|\hat{\phi}_{a^J}(\boldsymbol{\omega})\right|^2 + \frac{1}{2}\sum_{j \leq J}\sum_{r \in R}\left(\left|\hat{\psi}_{a^j,r}(\boldsymbol{\omega})\right|^2 + \left|\hat{\psi}_{a^j,r}(-\boldsymbol{\omega})\right|^2\right) \tag{9}$$

and ϵ small. Since the low-pass filter ϕ_{a^J} is normalized ($\int d\boldsymbol{x}\phi_{a^J}(\boldsymbol{x}) = 0$ or equivalently $\hat{\phi}_{a^J}(\omega = 0) = 1$), the above condition implies

$$\hat{\psi}_{a^j,r}(0) = 0 \tag{10}$$

or equivalently: all wavelets should average to zero. Furthermore, if we define a mother wavelet, called the Morlet wavelet, as follows

$$\psi(\boldsymbol{x}) = \mathcal{N} g_\sigma(\boldsymbol{x}) \left(e^{i\xi \boldsymbol{x}} - \kappa_\sigma\right) \tag{11}$$

where $g(\boldsymbol{x})_\sigma$ is a Gaussian and $\kappa_\sigma \ll 1$ has to be chosen to satisfy Equation 10, it will have the property that its Fourier transform is real, hence that if the input signal y is real, $U[j,r]y = U[j,-r]y$, allowing us to only consider positive rotations.

Instead of labeling a wavelet by its standard deviation in the spatial domain σ, it is in this case more insightful to label it by its bandwidth b in the Fourier domain, defined by

$$\hat{g}_\sigma\left(\pm\left(\frac{b}{2},0,0\right)\right) = \exp\left(-\frac{1}{2}\sigma^2\left(\frac{b}{2}\right)^2\right) = \frac{1}{\sqrt{2}} \tag{12}$$

leading to

$$b^2 = \frac{4\ln 2}{\sigma^2}. \tag{13}$$

The Fourier transform of the Morlet is

$$\hat{\psi}(\boldsymbol{\omega}) = \mathcal{N}\left(\hat{g}_\sigma(\boldsymbol{\omega} - \boldsymbol{\xi}) - \kappa_\sigma \hat{g}_\sigma(\boldsymbol{\omega})\right). \tag{14}$$

It is, apart from the small factor κ_σ, centered at $\boldsymbol{\xi} = (\xi,0,0)$ with bandwidth b. The requirement $\hat{\psi}(\mathbf{0}) = 0$ leads to

$$\kappa_\sigma = \frac{\hat{g}_\sigma(-\boldsymbol{\xi})}{\hat{g}_\sigma(\mathbf{0})}. \tag{15}$$

The dilated and scaled wavelet becomes in the Fourier domain

$$\hat{\psi}_{a^j,r}(\boldsymbol{\omega}) = a^{-dj}\mathcal{N}\left(\hat{g}_\sigma(a^j r^{-1}\boldsymbol{\omega} - \boldsymbol{\xi}) - \kappa_\sigma \hat{g}_\sigma(a^j r^{-1}\boldsymbol{\omega})\right) \tag{16}$$

so that it is (apart from the small corrective term κ_σ) centered at frequency $\boldsymbol{\omega_c} = a^{-j} r \boldsymbol{\xi}$ with bandwidth $b_{a^j} = a^{-j} b$. Note that the corrective factor κ_σ is invariant under dilation and rotation.

We choose the normalisation factor of the mother wavelet $\mathcal{N}$, in order to be able satisfy Equation 8, as the inverse of the maximum of the "raw" Littlewood-Paley sum (which excludes the contribution of the low-pass filter)

$$\mathcal{N}^{-1} = \max_{\boldsymbol{\omega}} \frac{1}{2} \sum_{j \le J} \sum_{r \in R} \left(\left|\hat{\psi}_{a^j,r}(\boldsymbol{\omega})\right|^2 + \left|\hat{\psi}_{a^j,r}(-\boldsymbol{\omega})\right|^2\right). \tag{17}$$

3.2 Discretizing $SO(3)$

We still need to specify which rotations we will use to build our filter bank, and, consequently, what the correct values of ξ and σ are to satisfy Equation 8. Thanks to the Morlet wavelet having a real Fourier transform and our CT images being real, we only have to choose a finite number of positive rotations so that our scaled and rotated wavelets cover (are non-zero) on as much of the frequencies

$$(\omega_x, \omega_y, \omega_z) \qquad 0 < \omega_x < \pi, -\pi < \omega_y, \omega_z < \pi \tag{18}$$

as possible, where π is the Nyquist frequency in radians (equal to $N/2$ for an $N \times N \times N$ image). In practice, we can impose the additional constraint

$$\omega_x^2 + \omega_y^2 + \omega_z^2 \le \pi^2 \tag{19}$$

since we don't care so much about the highest possible frequencies in the signal (they are very small anyway).

A scaled and rotated wavelet has Fourier support around $\boldsymbol{\omega_c} = a^{-j} r \boldsymbol{\xi}$ with bandwidth $b_{a^j} = a^{-j} b$. Hence, because $\boldsymbol{\xi} \propto (1,0,0)$ and we essentially want to cover the half-ball described by Equation 18 and Equation 19, we should choose our rotations r such that they map the unit vector $(1,0,0)$ onto n points on the hemisphere with north pole $(1,0,0)$ such that the pairwise distance in $\mathbb{R}^3$ is maximal (i.e. they are "evenly spread out" across the hemisphere).

Placing our points on the corners of a platonic solid is ideal, but only works for at most 10 points (the regular dodecahedron). For a solution for general n, I have used the Fibonacci lattice, which distributes n points to maximize their distance *along the sphere* in an approximately optimal way, but this should be a good approximation to the problem at hand when to number of points is large.

A Fast implementation

A.1 Downsampling in the Fourier domain

I downsample in the Fourier domain by a factor 2^j by simply setting to zero all frequencies outside the

range $[-\frac{\pi}{2^j}, \frac{\pi}{2^j}]$ in each dimension. This ideal low-pass filter corresponds to a convolution with a normalized sinc-filter in the spatial domain.

For a discrete-time signal $x[n]$, the Fourier transform is

$$X[\omega] = \sum_{n=-\infty}^{\infty} x[n]e^{-i\omega n}. \tag{20}$$

Downsampling the spatial signal x by a factor D, i.e.

$$x_D[n] = x[Dn] \tag{21}$$

corresponds in the Fourier domain to

$$X_D[\omega] = \frac{1}{D}\sum_{k=0}^{D-1} X(\frac{\omega - 2\pi k}{D}). \tag{22}$$