Bridging deep learning and explicit mathematical transforms for adaptive, interpretable machine learning

Motivation: Modern machine learning systems are data-hungry and hard to interpret

Deep neural networks (DNNs) today are widely used, from self-driving cars to speech recognition [1]. They have shown impressive predictive performance, which could provide massive benefit to society if applied to the right domains. However, DNNs are often restricted to idealistic settings because they (1) require enormous amounts of data and (2) are nearly impossible to interpret, preventing their use in critical domains such as medicine.

To mitigate these problems, I propose to bridge modern data-hungry DNN methods with rigorous mathematical transforms in the lab of my advisor Professor Bin Yu (Statistics/EECS) at UC Berkeley. To ground these methods, we will focus on two domains: neuroscience (by collaborating with the lab of Professor Jack Gallant at the UC Berkeley Dept. of Neuroscience /EECS) and materials science (by collaborating with the lab of Professor Gang-yu Liu at the UCD Dept. of Chemistry).

Review: The scattering transform - an effective, explicit representation for signals

The scattering transform, initially introduced by Stephane Mallat [2], produces useful representations of signals for many tasks. The scattering transform is defined as a series of wavelet transforms and low-pass filters. Formally, for a signal x, a low-pass filter ϕ, and a set of wavelets ψ, scattering coefficients (for three levels) are obtained as:

$$S_0(x) = x * \phi, \qquad S_1(x, \lambda_1) = \left|x * \psi_{\lambda_1}\right| * \psi, \qquad S_2(x, \lambda_1, \lambda_2) = \left|\left|x * \psi_{\lambda_1}\right| * \psi_{\lambda_2}\right| * \phi$$

In the end, these coefficients are aggregated over L levels: $Sx = (S_m x)_{0 \le m \le L}$. For a signal $f \in L^2(R^d)$ (such as images or audio recordings) the representation obtained by the scattering transform $\phi(f)$ fulfills two desirable properties:

1. Translation invariance: $\phi(f(x)) = \phi\big(f(x-c)\big)\ \forall f \in L^2(R^d), c \in R^d$
2. Lipschitz continuity: $\forall f, h\ \left|\left|\phi(f) - \phi(h)\right|\right| \le \left|\left|f - h\right|\right|$

Proposed research part I: Augmenting the scattering transform with DNN techniques

The scattering transform is intimately related to DNNs (in fact, it can be implemented as a DNN [3]). DNNs have had success by specifically learning peculiarities of particular datasets, but are relatively uninterpretable and data-hungry. On the other hand, the scattering transform contains useful general representations of signals, but may miss signals specific to particular datasets. As a result, the scattering transforms often (1) misses key filters and (2) has many unnecessary filters. Here, we aim to mitigate these problems by learning parameters for the scattering transform via DNN techniques.

(1) Adding key filters. The scattering transform consists of tunable parameters which can be adapted to suit specific datasets (e.g. the choice of wavelet basis, the number of levels of scattering, and the number of scattering orientations per level). Perhaps most important is the

wavelet basis, which forms an orthonormal basis for the signal: $\{\psi_{j,n}(t) = \frac{1}{\sqrt{2^j}}\psi(\frac{t-2^j n}{2^j})\}$.

Wavelet bases can be quite diverse while still fulfilling the desirable properties above. Here, we propose parameterizing the basis and learning parameters via backpropagation in the scattering transform (followed by a simple linear classifier), which is differentiable:

$$\theta = \max_{\theta} g_{acc}(\phi_{\{\psi_{j,n}(t;\theta)\}}(X))$$

Here, g_{acc} represents the accuracy of a classifier trained on the scattering representation $\phi_{\{\psi_{j,n}(t;\theta)\}}$ of a training data set *X,* where the representation is now parameterized by θ. As θ is learned, the wavelet basis becomes more and more suitable for the data set under examination while still maintaining a parametric form which can be rigorously interpreted. Additionally, it imposes a bias that can help algorithms learn in a data-limited setting.

This approach is both feasible and potentially impactful, as it relies upon bridging two well-established perspectives on generating representations: DNNs and explicit mathematical transforms. Such a parameterization is possible, and in fact parameterizations exist already for some general classes of wavelets (e.g. generalized Daubechies wavelets [4], generalized Coiflets [5]). These wavelet bases can be differentiated with respect to θ, and then θ can be learned via backpropagation, a well-established technique used to train DNNs. New work is required to (i) efficiently implement this learning process, following similar principles as the DNN implementation of the scattering transform [3] and (ii) develop a more general parameterized wavelet basis for different data sets, following similar principles as previous attempts [4-5].

(2) Removing unnecessary filters. The scattering transform contains a full representation of a signal, but for certain tasks this entire representation is not needed. To measure the usefulness of a filter for a task such as image classification, we use the intuitive CAR index introduced previously by the Yu lab [6]. New preliminary results on image classification tasks using this index suggest that the scattering transform provides many redundant features for classification. Using the CAR index, we can adaptively tune non-differentiable parameters of the scattering transform (such as the number of levels and number of orientations at each level). We can start with an overparameterization (large number of layers L, large number of orientations per layer $O_1, O_2, \dots, O_n$) and then solve the following nested optimization problem:

$$\max_{L,O_1,..O_n} \max_{\theta} g_{acc}(\phi_{\{\psi_{j,n}(t;\theta)\}}(X))$$

where $g_{acc}(\psi(X_{train}))$ is the accuracy of a model trained on a training dataset. The inner maximization problem can be solved via the method in part (I) and the outer maximization problem can be solved using the CAR index to automatically decide which filters are important and how to correspondingly change the outer parameters. To ensure the stability of the filters we remove, we propose to retrain the classifier multiple times with random restarts before removing a filter.

Extensions. Note that these two techniques can be extended for use in an online setting for adapting to changing conditions. Moreover, the proposed class of models could be extended beyond parameterizing the wavelet basis of the scattering transform to incorporate learned convolutional filters in new ways (for instance, parameterizing a set of constrained convolutional filters which is broader than the scattering transform).

Proposed research part II: Specific applications to problems in neuroscience and nanoscience

Investigating the visual system via interpretable, predictive models. In collaboration with Jack Gallant's lab, we aim to apply the above technique to interpret data from neuroscience. An experiment was conducted in the Gallant lab in which macaque monkeys were shown natural movies while neurons in visual area V4 of the monkeys were recorded from. The challenge is to build models that can predict the recorded neural activity from the movie data and then interpret these models to understand what the neurons respond to. Previous results using scattering transform representations have shown promising performance on predicting neural responses, but better representations are needed for these models to compete with DNN models. Changing the wavelet basis could allow for better predictive performance (particularly as it matches some theories of visual system function) and would provide more interpretable results. These results would allow for a better understanding of the visual system and possibly provide principles for how human sensory systems perform computations.

Characterizing predictive properties of nanomaterials from images. In collaboration with Gang-yu Liu's lab, we aim to use a small set of nanomaterial images to develop models for how different conditions affect the development of different properties in nanomaterials. If these models are both interpretable and predictive, they can help nanoscientists ascertain the physical effect of different procedures for generating materials, allowing them to discover procedures for developing new nanomaterials. The proposed procedure can narrow down a key set of features which can describe a nanomaterial texture and still predict certain physical properties well, despite having limited data. Adding new filters can yield better-predicting models and removing unnecessary filters yields more concise/accurate interpretations (which are required for a nanoscientist).

Deliverables. I plan to begin by implementing different classes of parametric wavelets and training them via backpropagation as described in section (1). Then, I plan to apply the methods described in section (1) and (2) to the above problems in neuroscience and materials science. After observing the learned parametric wavelet representations in these applications, I plan to derive some results on the stability of such representations. All implementations and algorithms will be released as open-source software.

Broader impacts

The ability to (i) build interpretable models and (ii) learn from small data, as is proposed here, can greatly advance scientific research. By building models of the visual system that can accurately fit recorded data, neuroscientists can understand fundamental aspects of how the brain processes information while maintaining the interpretability of their models. Moreover, this same procedure could be used in various scientific applications which require both predictive accuracy and interpretability of models, such as genomics. Additionally, learning from small data can allow the spread of machine learning in many new scientific fields such as nanoscience and particle physics, where data collection is prohibitively difficult or expensive. In nanoscience specifically, this method can help to discover the effects of various conditions on the creation of nanomaterials and how they can be created more effectively. Overall, these classes of models can help scientists to build better-predicting models while yielding new insights in a variety of domains by bridging insights from DNNs and mathematical transforms.

References

[1] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436.
[2] Mallat, S. (2012). Group invariant scattering. *Communications on Pure and Applied Mathematics*, *65*(10), 1331-1398.
[3] Bruna, J., & Mallat, S. (2013). Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1872-1886.
[4] Vonesch, C., Blu, T., & Unser, M. (2007). Generalized Daubechies wavelet families. *IEEE Transactions on Signal Processing*, *55*(9), 4415-4429.
[5] Wei, D., Bovik, A. C., & Evans, B. L. (1997, November). Generalized coiflets: a new family of orthonormal wavelets. In *Signals, Systems & Computers, 1997. Conference Record of the Thirty-First Asilomar Conference on* (Vol. 2, pp. 1259-1263). IEEE.
[6] Abbasi-Asl, R., & Yu, B. (2017). Structural Compression of Convolutional Neural Networks Based on Greedy Filter Pruning. *arXiv preprint arXiv:1705.07356*.