Cyberinfrastructure for science-driven analytics with limited data

Pavan Turaga
Arizona State University
Associate Professor in Electrical Engineering and Media Arts

Sensing modalities have been exploding in areas as diverse as materials science, biology, geographical and space science, with the resultant data as diverse as images, time-series, point-clouds, and functional-data. However, with the rise of such scientific datasets, analytics techniques to convert the raw data to actionable insights have lagged behind. This is because in many of these areas, it is time-consuming to provide detailed human annotation of key concepts, or labels, that can be used to train effective machine learning techniques. Further, in these applications, often the goal is not to perform simple tasks like classification, but to arrive at scientific insights. Due to these reasons, standard applications of neural networks to these domains has resulted in slower progress than anticipated.

For instance, in material science imaging data, for use in 4D material characterization, the imaged data is known to have characteristics at different spatial scales, very different from natural images. This makes standard application of techniques like image segmentation and classification difficult. Further, annotating materials datasets is a highly time-consuming process and acquiring new samples are also time consuming. Simple data augmentation methods are currently used to make visual analytics with deep-nets more robust but also fail as augmentation methods are limited to simple effects like rotations, affine transforms, noise, blur and other factors; increasing the training-set requirements and training-time. Their generalizability beyond the factors considered is often unknown in scientific analytics.

To overcome these limitations, new deep-learning architectures motivated by directly encoding physics-based constraints may help. These constraints include knowledge of imaging physics, illumination models, view-invariant representations, and invariance to image-quality degradation. Turaga work at the Geometric Media Lab uses methods rooted in geometry and topology to enforce these constraints analytically either in loss functions, or in constraints over latent spaces. This can lead to robust architectures providing higher performance under new imaging modalities especially under low-shot learning paradigms.

We believe that we are at an important juncture where interdisciplinary insights from scientific domains, machine learning, cyberinfrastructure can come together to develop a new class of flexible techniques that are:

- rooted in known domain science
- leverage new mathematics from geometry, topology, functional-analysis
- leverage existing cyberinfrastructure from other domains
- adaptable to small datasets
- provide robust, interpretable, and actionable scientific insights.

Geometric Media Lab