Research Statement

Jinen Setpal

Department of Electrical & Computer Engineering, Purdue University.

When I first ventured into machine learning, better data was stressed as the most straightforward way to improve model performance. Technically, this is true: multilayer perceptrons encapsulate the hypothesis space of modern feedforward architectures. *Practically*, however: it's unreasonable to expect multidimensional convolutions, or the attention mechanism to be intermedially obtained by backpropagation. By encoding intuitive biases, researchers developed architectures that establish state-of-the-art performances which aren't backpropagatable even with *perfectly* engineered data. I've began studying aspects of deep learning theory that enable encoding these biases and am particularly excited to explore encoding inutition to establish domain robustness through my doctoral program.

Formally, my primary research interest lies in Deep Learning Theory; specifically, **using theoretically-backed approaches to encourage distributional generalization**. I'm currently exploring *intrinsic interpretabitily*; the objective of which is to develop neural networks that are interpretable by design. By constraining model optimization to conform to high-level notions of interpretability, I aim to establish a performant, generalizable basis for learning. My hypothesis is that setting single objectives as multi-levelled optimization tasks should allow for a human-like interpretation of the challenge. This should discourage shortcut learning, transforming loss landscapes to those with greater convexity, thus increasing the likelihood and quality of convergence.

A concrete example of what a successful implementations of this approach looks like can be found in the following pre-print: `https://dagshub.com/jinensetpal/contrastive-optimization/src/main/paper/main.pdf`. What's encouraging is that the **exact same architecture** trained using the interpretable objective I derived learned weights that implicitly reduced cross entropy loss more than optimizing *directly* for cross entropy loss! In addition to being more performant, it is also more interpretable, generating a provable attention map of the image regions used for classification.

If you also find this interesting and would like to discuss this research direction further, please reach out: jinen@setpal.net.