N. Rich Nguyen

RESEARCH STATEMENT

With the emergence of Big Data, self-driving cars, and the Internet of Things in the last few years, machine learning methods have proven to be astonishingly effective for predictive and deep insights. My research interest lies in developing machine learning methods to understand, classify, and model objects in images and videos while reducing the training effort required by the human experts. To this end, I focus on transfer learning methods for classification, particularly the regularization of objective function, which learns both the structural and target-specific parameters simultaneously.

Past Research My past research can be broadly categorized in the following research theme: applications of machine learning in biological experiments. To accomplish the research objective, my doctoral research has centered around solving two problems. First, training the classifier without a sufficient number of samples could often lead to poor performance because it is susceptible to over-fitting. The number of available samples of a new biological class can be limited when the setting changes in image capturing equipment and imaging modalities. Second, the distribution of training data in biological images is usually unbalanced. Thus, a majority of training by the human experts is spent on searching for appropriate data to be labeled, which is often an expensive process.

Contributions My contributions have been spanned on three components: 1) Integrating relative entropy on the prior shape and size distribution of blood cells in partnership with Carolinas Healthcare Systems; 2) Utilizing spatial-temporal distance of unlabeled data to create pseudo-labels for training a detector for social insects; and 3) Developing active search approach to acquiring most informative training data for a new pollen data set in collaboration with University of Arizona. Since August 2010, I have co-authored seven papers in conferences in applications of computer vision (4), machine learning (1), and liver diseases (2). Additionally, I have also co-authored three articles in respected journals in biomedical engineering (1), physiology (1), and trauma (1). Since then, my research has stimulated a good amount of follow-up work (70 citations on Google Scholar).

Future Directions In the upcoming years, my research agenda will be strongly influenced by an array of collaborative, data-driven, large-scale, and sustained efforts that emerge in the advances of Computer Science (CS) Education. NSF especially encourages the tools and models for teaching and learning aimed at promoting success and inclusion in computing across diverse populations, particularly those populations that have been traditionally underrepresented in CS and STEM fields. Despite being present across all disciplines, the issues of attrition and retention can be most noticeably seen in CS. To help CS retention, I intend to develop classification models trained on a combination of academic performance, student engagement, and demographic data to show that exploring these three aspects possibly resulted in early predictions that are far more accurate and actionable than existing models. Early warning indicator systems that harness student-level data to provide educators with early flags for students who may be at risk are becoming increasingly popular. I am curious to know which predictive and preventive measures would be the potential success of efforts to broaden participation in CS at the undergraduate level. Specifically, my research agenda decomposes into three potential projects, which I briefly describe below:

1. **Transfer Learning in Deep Convolutional Neural Network (CNN) with Limited Educational Labeled Data** Tremendous progress has been made in object recognition with deep CNN. However, a large-scale dataset is indispensable when training a deep CNN, such as ImageNet that contains about 22,000 classes and nearly 15 million labeled images. Unfortunately, there exists no large-scale annotated academic domain comparable to ImageNet, as data acquisition is challenging and quality annotation is time-consuming. Therefore, the limited labeled on educational data becomes an obstacle to train a deep CNN. To solve this problem, I plan to expand my doctoral research to making knowledge learned from sufficient unlabeled data transferable to target data.

2. **Machine Learning Approach to Predict and Prevent At-Risk Students from Unfavorable Learning Outcomes** I aim to expand the functionality of learning-based models to allow for a careful prioritization of the students thought to be at risk. Using learning analytics to investigate patterns of participation in extra-curricular activities, I want to quantify student engagement on a large-scale and using it to identify students that may be leaving a STEM discipline, even when they are not displaying poor academic performance.
3. **Data-driven Affective Models to Broaden CS Participation** One of the biggest challenges in broadening participation in CS is how to integrate CS and computational thinking into the current K-12 curriculum effectively. Many teachers are willing to learn, but finds themselves lack (1) relevant materials and (2) a supportive community. I propose affective learning models using large-scale data set for information sharing to foster a sense of belonging among the underrepresented students in CS. To this end, I intend to launch CodeNC.org, a teacher-empowered program to promote CS in K-12 teachers across North Carolina.

I plan to conduct research that makes an impact on the educational experience at the undergraduate level. I aim to build relationships to facilitate research collaboration with faculty and students at historically underrepresented institutions and to pursue internal and external funding to support this type of collaboration. My research has been informed by a belief that machine learning techniques must have relevance to solve societal problems. The NSF Directorates for Computer and Information Science and Engineering (CISE) is interested in receiving proposals that advance the CS for All Initiative. Thus, I intend to write grant proposals to obtain funding from CISE and CSForAll to solve the above problems, apply my expertise to learn interesting educational trends, and foster collaborations with other faculty members in the department.