S+T

NARRATIVE

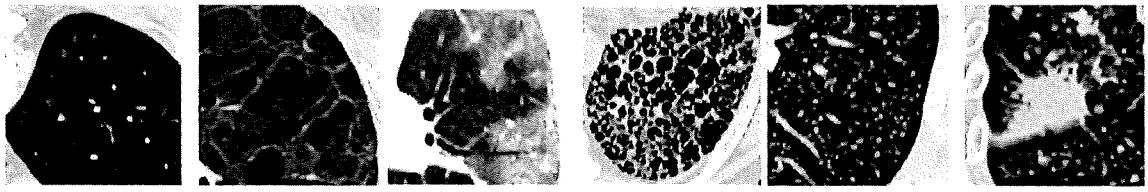
Project Director: [redacted]

Objectives

Chest computed tomography (CT) scans are widely used for automatic detection and classification of pulmonary diseases. CT scans are three-dimensional (3D) image datasets that capture properties of tissues and organs in the body. Manual interpretation of a large number of these scans by radiologists is time-consuming and can be error-prone, especially when healthcare professionals are carrying a heavy workload. For this reason, there is considerable interest in developing computer-aided diagnostic (CAD) systems that can screen and/or detect pulmonary diseases automatically, efficiently and with reduced risk of detection errors, which in turn can help radiologists optimize their diagnostic decisions. The goal of this project is to develop a CAD technique for automatically classifying whether lungs are normal or infected by Interstitial lung disease (ILD). ILD represents a group of more than 150 disorders of the lung parenchyma. These infections typically manifest themselves as textured patterns in CT scans (Figure 1) and so, historically, *machine learning* techniques have been used to distinguish between patterns belonging to different diseases.

What is machine learning? It is a powerful algorithmic framework that "learns" the relationship between some "training data" and corresponding "labels". For example, imagine that we want to teach a computer to determine if a photograph contains a cat. In machine learning, one starts with a large "training data" image set, about half of which contain a cat and the other half of which do not. "Labels" can be denoted by "yes" or "no" to indicate the presence or absence of a cat or it can be more sophisticated such as boundary around the face of the cat. The training (or learning) process first extracts certain *features* (e.g. shape, color, texture, etc.) from these images that may help in differentiating cats to other things and then employs *classification* algorithms to build a relationship between these features and the "yes" or "no" labels. At the end of the training, the computer acquires a "model" of a cat so that when presented with a *new* photograph, it predicts whether the image contains a cat or not.

Traditionally, machine learning techniques employ image *features* handcrafted by experts to detect certain textural patterns. Recently, a new subfield of machine learning called *deep learning* has emerged that, instead of expertly crafting these features, *infers* them automatically by modeling high-level abstractions in data using multiple processing layers [1]. These layers are typically represented using Convolutional Neural Networks (CNNs), which have outperformed the traditional machine learning models for object classification in *natural* images, such as those of people, trees, animals, and buildings. This proposal will investigate usage of CNNs for *medical* image classification. This extension from *natural* to *medical* images is not trivial since CNNs require a variety and quantity of annotated image data generally unavailable from medical sources for research use. Consequently, this proposal will investigate: i) "transfer learning": an approach that employs existing networks trained on *natural* images and fine-tunes them for *medical* images[2], ii) novel CNN architectures suitable for classifying ILD in CT. The proposed method will be tested on publicly available databases containing 128 CT scans with ILD.



(a) Healthy (b) Emphysema (c)Ground glass (d) fibrosis (e) micronodules (f) consolidation

Figure 1: Visual aspects of the most common lung tissue patterns in CT of patients with ILDs. Note the different textured patterns.

Significance

This proposal will open the door for applying cutting-edge *deep learning* algorithms for medical imaging applications. Previous proposal in this topic was recently funded through the *Koret* project and it investigated the efficacy of *traditional* machine learning techniques. This proposal is a step towards investigating efficacy of *deep learning* algorithms in medical domain. Due to recent resurgence of *deep learning* and the obvious importance of medical imaging applications in improving patient healthcare (diagnosis, treatment, therapy, etc.), the combination of two has become an active area of current research. This proposal will develop computational framework to develop and implement deep learning algorithms in the medical domain. In addition, it will pave the way for the final frontier in CT based classification - that of employing the entire 3D data (as opposed to 2D slices) to aid in detection and classification of diseases.

This proposal will also provide PI with data that can be used in a collaborative NSF REU proposal from Sonoma State, which seeks to introduce a machine learning based research environment for undergraduates in Northern California, specifically to provide exposure in two key areas: i) deep learning algorithms and ii) their application in multiple STEM disciplines such as computer science, geology, biology and environmental enquiry. PI plans to submit this proposal to NSF's Division of Information and Intelligent Systems in August 2019.

Plan of Work

The primary goal of this proposal is software development. All the pertinent software libraries and image data has already been acquired by the PI. A workstation fitted with 2 Graphical Processing Units has recently been purchased. This proposal seeks an additional GPU to further increase the efficiency of this workstation. The software will be developed in several stages as outlined below:

1. Implement "transfer learning" based CNN and optimize performance by tuning several parameters. **(Summer 2018)**
2. Design and train a *new* Convolutional Neural Network that specifically identifies low-level textural features of the lung tissue. **(Fall 2018)**
3. Evaluate impact of dimensionality of input data; That is, whether the input to CNN is an image patch (2D), entire image slice (2D), or multiple image slices (2.5D). **(Spring 2019)**

Student Involvement

At least 2 undergraduate students will be selected through a competitive recruitment process to work on this project. The recruitment process will give equal weightage to coding experience and passion in working on this project demonstrated via a statement of interest. In addition to implementing the tasks detailed in the section above, students will be trained to employ software development and data management techniques that will prove essential in their future careers in academia/industry. These techniques include i) Testing the correctness of their code (on a smaller "toy" problem), ii) Catching bugs in code using IDE debuggers (such as PyCharm), and iii) Maintaining code through version control system (such as Github).

Student progress will be discussed through bi-weekly meetings during regular semester and weekly meetings during summer. At the end of each semester, students will present their work at several avenues: i) Computer science colloquium series (Nov 2018), ii) CSUPERB annual symposium (Jan 2019) and iii) SSU Research Symposium (May 2019).

References

[1] LeCun, Y., Bengio, Y., and Hinton, G., "Deep learning," Nature 521(7553), 436–444 (2015).
[2] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks?. Neural Information Processing Systems 27 (NIPS '14), pages 3320 - 3328