Research Projects

Overview

Machine learning has become a central technique in using computers to solve practical problems. It is used in web search, spam filters, image tagging and recognition, interactive voice activated programs like Apple's Siri, software that beats human professionals in chess, Jeopardy, Go, etc.

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. 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.

Recently, inspired by human brain architecture, a new subfield of machine learning called *Deep Learning* has emerged [1] and its applications are in diverse fields of medicine, finance, weather, earth sciences, etc. [2]. Instead of using hand-tuned features designed by experts, it *infers* features automatically from training data [1]. This is possible by learning from deep, layered and hierarchical models of data, typically represented using Convolutional Neural Networks (CNN) (Figure 1). CNN models require tuning of *millions* of parameters, which is largely possible due to recent hardware advances that allow parallel processing on Graphical Processing Units (GPUs) and due to the availability of large amounts of training data and labels through datasets such as ImageNet [3], which contains more than 1.2 million natural images of 1000+ object categories. But what happens if one does not possess GPUs or if one wants to classify images that are not among the 1000 object categories? An approach called "Transfer Learning" [4] can be employed, where features developed in *existing* (pre-trained) networks (on image databases like ImageNet [3]) can be used for classifying *new* data. This approach can be implemented on CPUs and doesn't require expensive GPUs.

Transfer Learning typically involves removing the last classification layer of the original (pre-trained) network and retrofitting it with a new classifier that is trained to classify new set of images. Another strategy, called "fine-tuning", involves extending this approach to not only replace and retrain the last layer but to continue adjusting the weights of the neurons in previous layers [5].

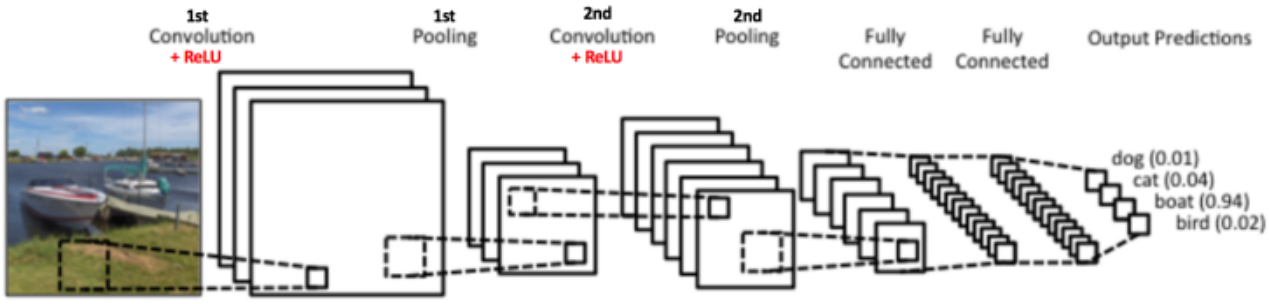


Figure 1: Example of a Convolution Neural Network (CNN) architecture showing 2D image as input into the network (leftmost) and a classification label (car, boat, dog, etc.) as output of the network (rightmost). Figure credit: http://tinyurl.com/zjjltev

Based on these techniques, following are some of the projects that will involve undergraduate students.

1. Geosciences: Towards building a cyber-infrastructure

Goal is this project is to automatically extract and label patterns from geological imagery, e.g., photomicrographs, SEM images, EBSD maps, EDS maps, microprobe imagery, aerial imagery, outcrop photos, etc. See Figure 2 for an overview.

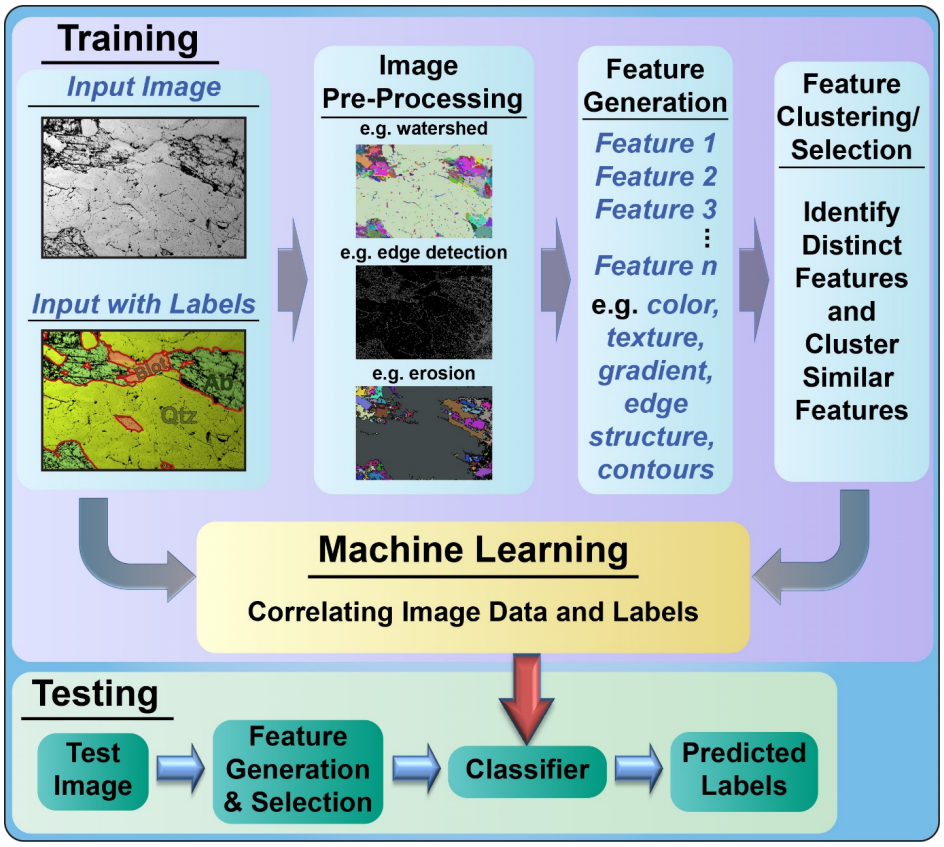
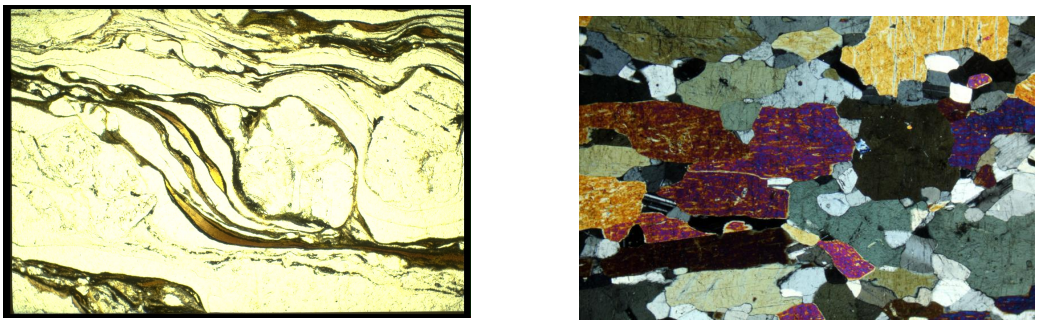


Figure 2. An example workflow for utilizing machine learning for automatically labelling geological imagery (e.g., minerals, microstructures, etc.). The "Training" module identifies features and then employs machine learning algorithms to learn the relationship between the input image and the corresponding labels. The "Testing" module uses the learned classifier to label a new test image. By utilizing "Deep Learning" [1] the feature identification component is automated.

Initial experiments based on CNNs have shown some success in identifying photomicrograph images (Figure 3) that either have sigma clasts or not.



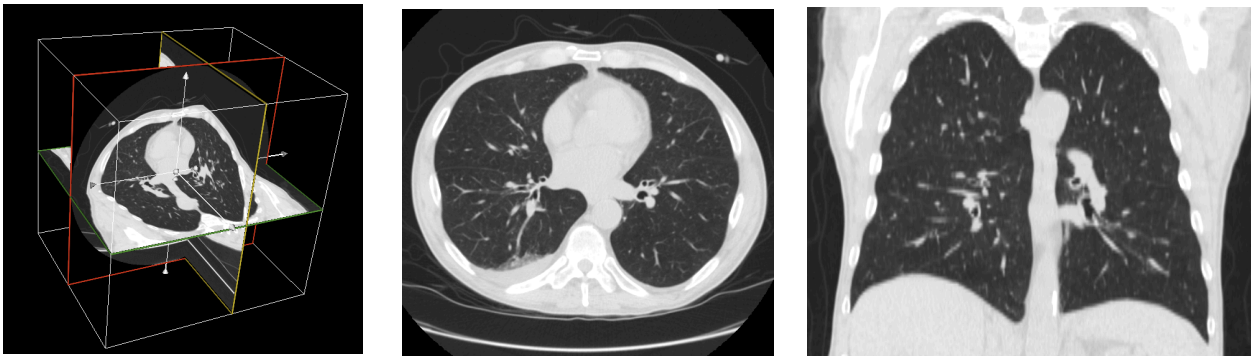
(a) (b)

Figure 3: Photomicrographs showing presence or absence of sigma clasts. a) Present, b) Absent

2. Medical Imaging (Lung image analysis): Classification of Interstitial Lung Disease in Chest Computed Tomography Scans

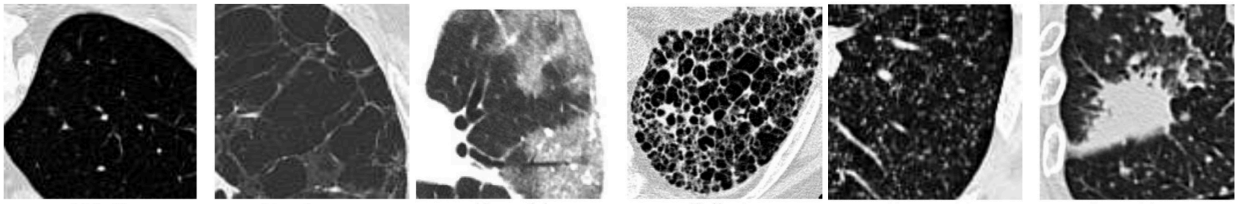
Chest computed tomography (CT) scans are widely used for automatic detection and classification of pulmonary diseases [6]. CT scans are three-dimensional (3D) image datasets that capture properties of tissues and organs in the body (Figure 4). Manual interpretation of a large number of these scans by 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 [1]. These infections typically manifest themselves as textured patterns in CT scans (Figure 5) and so, historically, *machine learning* techniques have been used to distinguish between patterns belonging to different diseases [7].

This project will investigate efficacy of employing CNNs in the following ways: i) Using "Transfer Learning"/ "Fine-tuning" approach to overcome problem of limited annotated medical data and ii) Training the CNN "from scratch": an approach that designs novel CNN architecture and gets around the problem of limited data through a technique called "data augmentation" (generating new data through variations of existing data). Both methods will be tested on publicly available database containing 128 CT scans with ILD [8].



(a) (b) (c)

Figure 4: Example of a chest CT scan. a) 3 cross-sections showing the 3D nature of CT, b) Single axial slice of a CT scan (red plane in a) and, c) Single sagittal slice of a CT scan (green plane in a)



(a) (b) (c) (d) (e) (f)

Figure 5: Visual aspects of the most common lung tissue patterns in CT of patients with ILDs. Note the different textured patterns. Labels for image patches are provided along with CT data [8]. (a) Healthy; (b) emphysema; (c) ground glass; (d) fibrosis; (e) micronodules; (f) consolidation.

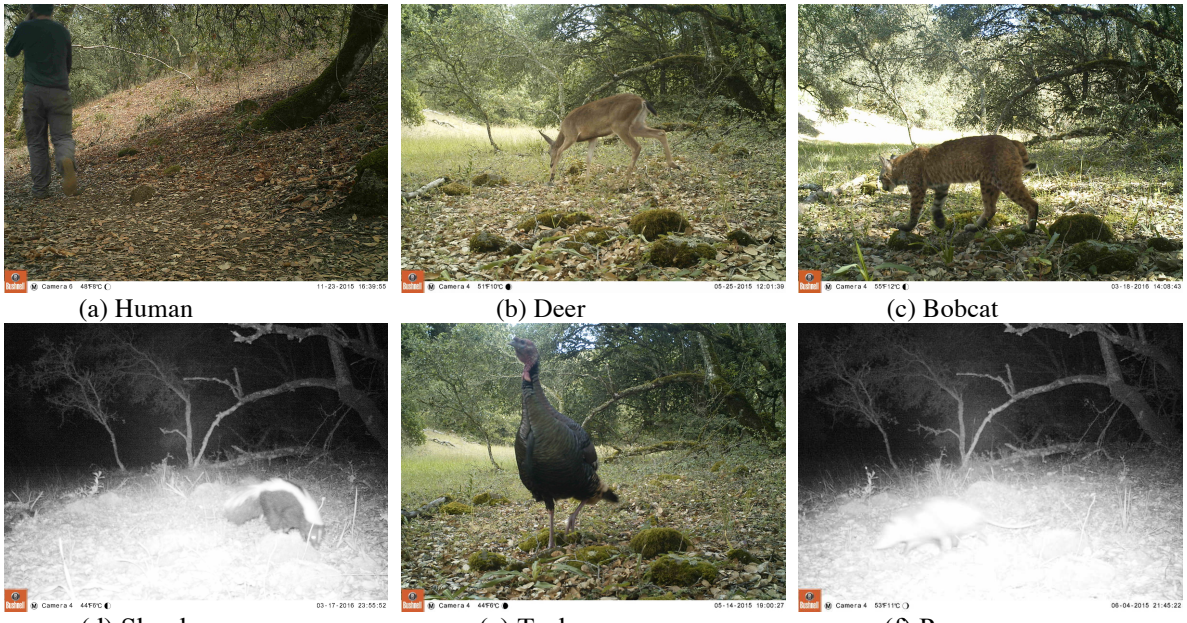
3. Natural scene classification: Classifying wild animals in images taken from cameras placed along trails in wild life preserves.

Automatic classification of wild animals through trail cameras can be helpful to study movement patterns of animals, type of animals in a specific region, inter-dependencies between different animal species., etc. These studies become important for taking appropriate measures when, for example, animal habitat is disturbed through human projects such as placing of power lines or due to construction of a highway. It will open the door to address more complex challenges involving animal habitat and movement in the wild.

The classification problem poses whether the animal is deer, mountain lion, squirrel, skunk, possum, etc. (Figure 6). The complexity of the problem arises from the unpredictable nature of wild life. Trail cameras take a picture whenever the motion sensors are activated. The location and size of the animal in the picture, thus varies a lot. Moreover, often the animal is only partially visible. Lastly, pictures taken at night (through an infra-red camera) present an added challenge due to their different appearance and due to pictures sometimes getting "white-washed" with camera flash when the animal is too close to the camera.

This project seeks to apply CNNs on a large dataset comprising thousands of trail images obtained from SSU's three preserves: the Galbreath Wildlands Preserve, Fairfield Osborn Preserve, and Los Guillicos Preserve. Goal is to assess performance of CNNs on several cohorts:

- Camera specific: Train and test on images from *same* trail camera. This exploits the fact that the size/location/viewpoint of animals seen from a single camera will be similar and so the algorithm can learn a more accurate model. The flip side is that at evaluation time separate models will need to be applied to images from corresponding cameras.
- Camera independent: Train and test on images from *all* cameras. This is the most general case where algorithm learns from the wide variety of images presented from all cameras. At evaluation time, a single trained model can be used for classifying any trail image.
- Camera contrary: Train on images from *one* camera, test on *another*. This case examines how well the algorithm can classify animals in a previously unseen location and viewpoint. It tests the robustness of the deep learning framework in modeling generic attributes of animals.



(a) Human (b) Deer (c) Bobcat

(d) Skunk (e) Turkey (f) Possum

Figure 6: Example of images of animals obtained from trail cameras placed at SSU preserves.

References

[1] LeCun, Y., Bengio, Y., and Hinton, G., “Deep learning,” Nature 521(7553), 436–444 (2015).
[2] https://www.nextplatform.com/2016/09/14/next-wave-deep-learning-applications/
[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in IEEE CVPR, 2009.
[4] 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
[5] http://cs231n.github.io/transfer-learning/
[6] U. Bag˘cı, M. Bray, J. Caban, J. Yao, and D. J. Mollura. Computer-assisted detection of infectious lung diseases: a review. CMIG, 36(1):72–84, 2012.
[7] Y. Song, W. Cai, Y. Zhou, and D. Feng, “Feature-based image patch approximation for lung tissue classification,” IEEE Trans. Med. Imag., vol. 32, no. 4, pp. 797–808, Apr. 2013.
[8] Depeursinge A. et al., Building a reference multimedia database for interstitial lung diseases. Computerized Medical Imaging and Graphics 36(3) pp 227-38, July 2012