Nate Kaiser
Adam Pollack
Chainatee Tanakulrungson

EECS349 Status Report

5/17/2017

Task

Our project goal is to predict the presence of lung cancer nodules (binary output) given small (e.g. 50 x 50 pixel) snippets from medical images labeled either “cancer nodule” or “not a cancer nodule”. This problem is unique and exciting in that it has impactful and direct implications for the future of healthcare, machine learning applications in personal data and decisions, and computer vision in general. The medical field is a very likely place for machine learning to thrive as medical regulations continue to allow increased sharing of anonymized data for the sake of better care. Not only that, but the field is still new enough that our project can mimic methods at the forefront of technology.

Data

Medical image databases are readily available, and millions of labeled images are freely available for download (many are even split into training and test sets already). We are using the lung cancer collection from the LUng Cancer Nodule Anlysis (LUNA) dataset at https://luna16.grand-challenge.org/data/. The data consists of thousands of annotated lung cancer nodules, which include labels as to whether each is “definitely a nodule” or “potentially a nodule”, and the (x, y, z) location in the 3D image set of each. We have used this information to create nodule (positive) instances and random location (negative) instances in order to train, validate, and test our learner. These sets will be randomly selected from one large pool of images, to keep selection bias from affecting the project.

Preliminary Results

Since this task is very complex and quite a challenge to implement from scratch, we have not yet run our image dataset through our planned learning algorithm (neural networks). However, we have preprocessed the data (learning a lot about CNNs and the requisite image processing along the way), installed and verified the necessary software (Python, TensorFlow, etc) on our machines, and have trained other neural networks (AlexNet, etc) on some already-preprocessed sample training sets on readily available data such as CIFAR10 (https://www.cs.toronto.edu/~kriz/cifar.html). Our initial runs with these practice datasets have taken a long time to complete (normal for deep networks learning images), so we have invested much of our time searching for resources available to Northwestern students for high-end GPU calculations, which we have recently secured. This has been one of the main challenges of the project, and we hope it will ensure quick training and testing times so we can try multiple iterations of hyperparameters.

Future Plans

At this point, we are close to having our dataset built, and once complete we will start training our learner. This will most likely be a challenge, since TensorFlow is very complex and can behave differently across different underlying hardware. Ideally, we will see promising results with our first attempt at tuning hyperparameters and go from there, but we expect and are fully prepared to iterate until we start seeing “correct” looking results. Some other outlying concerns we have:

- CNNs will require too much data or time to train and we will have to revert to feature-based learning techniques
- Training times will be too long in order to properly tune our hyperparameters and we will end up with lower than desired test accuracies
- Results will return unfavorable and we won't be able to explain why (since neural networks aren't easy to troubleshoot, this is a common concern)
- None of us has experience tuning large, complex networks like a CNN, so parameter tuning might be difficult or impossible
- There may not be time to try other models and compare the results (SVM, k-NN, etc)

These concerns aside, we are all very interested in machine learning for vision applications and are excited for the challenges of this project. While there may be a lot of work ahead of us, we plan on working toward satisfactory results, and learning a ton along the way!