In the second year of the project, we focused on three tasks: (1) stabilize the computing environment and resolve the remaining computer networking issues; (2) optimize the workflow for scanning and annotation so that we can consistently scan and annotate 2,000 slides per week, and (3) develop a baseline deep learning system that is able to handle data at scale. We refer the reader to the book chapter included in this annual report. This contains a very detailed summary of the project to date and represents a very nice attempt to document all the issues we have faced and resolved regarding the development of the hardware, data and machine learning technology.

1. COMPUTER NETWORKING ISSUES

The computing architecture developed for this project is more complex than one might expect because it must span multiple networks within Temple University. The attached network diagram provides an overview of the physical implementation of the systems (Figure 2). Although cloud computing is an enticing technology for many problem spaces, it does not come without its drawbacks, including, most notably, data storage costs and privacy. For the pathology corpus, it was essential that the operations performed on the data, particularly viewing and annotation, did not experience low throughput and high latency. Another essential requirement for the infrastructure housing the corpus, especially in clinical settings, was its robustness. This was because the data stored in the corpus would have to be HIPAA protected for clinical or diagnostic purposes.

For the storage architecture of the corpus, a robust and extensible distributed storage device was devised which employed numerous open source tools for the creation of a single filesystem. These machines were not allowed to sit on the same physical network due to concerns about the impact they would have on hospital operations. Creating the ability for these systems to communicate with one another in a secure but transparent way without burdening pathologists with complex VPN interfaces was no easy feat, requiring modification of several router tables and firewalls.

2. IMAGE SCANNING AND ANNOTATION

In digital pathology, an image of an analog specimen is captured as a whole slide image (WSI) using a laser scanner that produces an image with a resolution of 0.2 μm/pixel. Though a large number of slides can often be scanned automatically, approximately 5% of our slides require selecting manual focus points to help the scanner properly focus. WSIs can contain one or more specimens, which further complicates the machine learning problem.

We are using a Lecia Biosystems Aperio AT2 high volume scanner to digitize our slides. This scanner includes an autoloader consisting of 10 trays that can hold 40 slides each, resulting in a total capacity of 400 slides. Pre-scan snapshots must be taken before the scanner is set to perform full scans overnight, prohibiting complete automation of the scanning operation. This process takes around two hours for the 400 slides loaded. There are some cases where the scanner makes errors in its placement of the focus points and in determining the area of the scanning region. In such cases, the focus points must be manually placed. This makes the pre-scan snapshot phase of the scanning procedure labor intensive. Among the 400 slides set to scan overnight, we find about 2% are likely to experience a scanning failure. However, this number varies according to the quality of the stain applied to the slides. These failed slides are reviewed, readjusted, and scanned again the next morning. We have developed a fairly robust process for ensuring that we keep the scanner busy and in a typical week we can scan 2,000 slides.

The images and reports are stored on the file server using a file naming convention that is best explained by the table shown in the attached figure (Figure 3). The file naming convention is designed so that every file in the corpus has a unique filename, and simple UNIX commands can be used to locate data.

Using a single Leica Biosystems Aperio AT2 scanner, we are able to scan about 2,000 slides per week with a small team of undergraduate student workers. We have currently scanned over 60,000 slides. The statistics for the first 20,000 of these slides, which is available as a pilot corpus, are show in the attached figure (Figure 4).

The Leica Web Viewer is the annotation tool that pathologists are using to annotate images. It provides an assortment of shapes such as rectangles, circles, and ellipses that can be used to identify a region of interest. The SVS files created and stored in the petabyte server can also be viewed and annotated using ImageScope, which is free, proprietary software provided by Leica Biosystems. After an image has been annotated, an XML file is generated which contains the annotation details.

3. Deep Learning Experiments

The fundamental motivation behind developing the TUDP Corpus is to provide adequate amounts of data for training sophisticated deep learning systems.

For our initial baseline system development, we selected a preliminary dataset of 1,000 pathology slides averaging around $5,000 x 2,500$ pixels in size. These slides were selected based on an initial screening process which determined whether a mark from a stray marker (a grease pen) existed on the pathology slide. These visible marks served as the event to be classified. Every $50 x 50$ pixel patch was annotated by a team of nine annotators. Each image was classified as having a mark if at least 2% of a given patch contained a mark artifact. Of these 1,000 pathology slides, 500 slides were identified as having one or more marks while the other 500 did not have any marks.

For the deep learning system, we leveraged our work on EEGs. Perhaps the single biggest challenge in processing these high-resolution images is that they must be segmented into small sections as the entire image will not fit into memory. The segmentation of these images involved using a frame and window size to partition the images. In our baseline architecture there are a total of five convolutional layers along with a fully connected hidden-layer, and an output layer. All of the convolution layers use a Rectified Linear Unit (ReLU) activation function.

Scoring of the deep learning system was done using both a frame-level and a whole-slide methodology. The mean sensitivity rate for the mark class was 99.3% on WSI classification. On the evaluation set, our system obtained 99.4% for the mark class and 99.0% sensitivity for the null class. For the same set, the unprocessed frame-level predictions had a mean sensitivity of 99.37% for the mark class and 99.73% for the null class. Error analysis shows that the postprocessor, as expected, rejected images with very small marks. Performance was sufficiently high that we were ready to move to the next step, which is to establish baseline performance on a small, realistic corpus of breast cancer slides. This is under development currently and should be completed by the end of the second year of this project.