Recognizing Patterns of Cancer in Histology Imagery Using Deep Learning

Ted Hromadka¹, LCDR Niels Olson² MD, LT Daniel Ward² MD, CDR Arash Mohtashamian² MD, Ken Abeloe¹

¹Integrity Applications Incorporated℠, ²US Navy NMCSD

Presented at GTC 2016
Background – prostate cancer is a significant problem

- US military’s hospitals care for disproportionately more male patients
- Prostate cancer is second-leading cause of cancer death in American men
  - Approximately 220,000 new cases per year

- Early screening involves a blood test for prostate-specific antigen (PSA) or a digital rectal exam (DRE)
  - If those tests generate abnormal results, then a prostate biopsy may be required

http://www.va.gov/vetdata/docs/quickfacts/Population_slideshow.pdf
http://www.cancer.org/cancer/prostatecancer/detailedguide/prostate-cancer-key-statistics
Each biopsy procedure creates around 12 samples

- Prostate biopsy is conducted by taking “core samples” using a hollow needle
- After processing, 5 micron sections of these samples are placed on glass slides, stained, and manually interpreted by a pathologist under a microscope.
Analysis is very labor-intensive

- Digital scans are opened with custom viewing software from the microscope vendor
  - Multiple zoom levels available up to 40x. This dataset was scanned at 20x.

- Pathologist will annotate cancerous regions with polygons drawn by hand with a mouse

- Process requires careful judgment and is susceptible to fatigue and stress factors. Polygons cannot be edited once drawn (e.g., at higher magnification).
Biopsy analysis is challenging

- Tissues can be difficult to differentiate
- Cancerous region may be only partially sampled by the needle
- This is an image classification problem
Apply deep learning techniques to this image classification problem

- IAI was using Caffe for ship detection and classification in maritime aerial imagery

- Believed NVIDIA’s DIGITS software offered promising approach for the histology problem
Deep learning in a nutshell

- GPU-enabled evolution of artificial neural networks from 1990s
- Each layer is a set of “neurons” with weighted connections
- Each neuron responds to its unique aspect of the input data with varying degrees of strength
- Different weights compute different functions
- Training the network “teaches” it a complicated function
  - Supervised vs unsupervised learning
- Modern computing hardware allows more layers of neurons... “deep” learning
  - Reinforcement learning
- Several open, GPU-enabled frameworks (Caffe, Torch, Theano, DL4J, TensorFlow)
- Convolutional neural networks excel at image recognition
Puppy or bagel?
Specifications

- Imagery
  - 202 annotated full-size color SVS images → 106,024 image chips
    - Average full size image ~ 845 MB
  - Annotated by Navy pathologists

- System
  - NVIDIA GeForce GTX980 GPU (single card) via Intel Haswell-E PCIe 3.0
    - Maxwell architecture, 2048 CUDA cores, 4GB memory, NV driver 352.63
  - 6-core Intel Xeon E5-2603 v3 at 1.60 GHz with 16GB DDR4
  - Ubuntu 14.04, DIGITS 3.0-rc3, CUDA 7.5, cuDNN v4, NVCaffe 0.14
Used MATLAB image chipper to prepare the images

- Split SVS into image chips of size 256x256 pixels at the 4:1 zoom level
- Chipper labels each image chip based on XML annotation polygons (50% inclusion rule)
- Chipper 2.0 also used pixel averaging and histograms to determine if chip was a “blank” or an “ink” smear

[XML parser built on work by Andrew Janowczyk](http://www.andrewjanowczyk.com/)
Naïve results were terrible

- Simple “cancer / not-cancer” labeling was a disaster
- Immediate 50% accuracy for a binary classifier meant that it was just a random guess
Solution: refine the training categories

- Bad data (blank areas, ink marks)
- More tissue types (fat)
- Manually inspect the input data for anomalies

- Still using stock GoogLeNet network
- Additional training epochs had minimal effect
Cancer or not cancer?
5 categories of refined training data => raised accuracy to 90%
How accurate is the measure of accuracy?

- Elmore et al – Breast Biopsy Concordance study found only **75% agreement** between expert pathologists

- Need protocol for the confidence levels
  - What threshold to use when network gives it a substantial chance of cancer?
In progress - adding more categories to improve accuracy

- Seminal vesicles
- Lymphocytes
- Corpora amylacea
- Blood
- Nerves
- Muscle (healthy)
- Stroma
- Gleason scale
- Perineural invasion
- Atrophic glands
- Atrophic prostate necrosis
In progress – looking for ways to handle pragmatic labeling

- Training data suffers from inaccuracies
  - Annotation was not meant for training neural networks
  - Not pixel-perfect
- Artifacts due to the scanner or tissue preparation
  - Striping
  - Ink
- Experimenting with statistical solutions to noisy data
Project assessment: bulk of time was spent on data preparation

- MATLAB time mostly spent moving data
- DIGITS greatly facilitated the DL training
- Labeling images

- Annotate images
- Write MATLAB chipper
- Run MATLAB chipper on data set
- Install & configure DIGITS
- DIGITS - create database
- DIGITS - train 1 network
- DIGITS - run 1 chip on network
- Caffe - run 1 full image on DNN
Automated image classification step is 50% faster than a pathologist

- Chipper, classifier, output rendering = 29 minutes, vs “less than an hour” for a pathologist
- Still needs a pathologist to review the output for final determination
- Will be faster on better hardware
- Data transport is a bottleneck to using HPC assets, but not an impossibility
  - Upload raw microscope image to Navy DSRC
  - Run image processing on those GPU nodes
  - HPCMP Portal “Virtual App” for final pathologist image review
- Also considering Google/AWS/Azure services deployment, but HIPAA complications
Next steps – fully automated process

- No signs of overfitting – seek more data

- Try 128x128 chips to reduce chance of multiple tissue types per image

- Software pipeline
  - Digitization scan > Chipper > DL Classifier > Heat Map > Viewer
Other approaches

• Similar work ongoing by Hao Chen at Chinese University of Hong Kong
• Multispectral imaging
• ImageJ – Cell morphology
• Andy Beck’s lab at Harvard
• Antonio Criminisi and Steve White with glioblastoma: WhiteMarshLabs
• Aperio GENIE (uses older machine learning algorithms, entirely CPU-based)
Conclusions

• Vast majority of time spent on preparing the data

• Unsupervised learning is the future

• Works surprisingly well