Accelerating Image Analysis Workflows using Deep Learning

Yanling Liu
Manager, Imaging and Visualization Group, ABCS/BIDS, FNLCR

February 21, 2018
Overview of the Imaging Group

Focused on the practice of computational science as it applies to problems in cancer and biomedical research
Turning Image to Knowledge to help/drive/accelerate/advance cancer and biomedical research

**Image data**
- Annotation
- Metadata
- Radiology
- Microscopy
- Pathology
- EM
- Multi-spectrum
- etc.

**Analysis**
- Ground Truth generation/validation
- Quantification (based on segmentation)
- Registration (intra-modality and cross-modality)
- Feature extraction and measurement (vessel network analysis, for example)
- Algorithm development
- Machine learning and Deep learning
- etc.

**Knowledge to help domain experts**
- Prescreening
- Diagnosis
- Treatment
- Precision Medicine
- etc.

**Support**
- Storage/Network
  - Database, Archive/PACs
- CPU, GPU, Clusters, HPC
- Visualization, HCI, VR/AR
  - Web/Mobile access, Database, etc.
Goals for Imaging and Visualization Group
Impact on cancer and biomedical research at FNLCR/NCI/NIH

• Reproducibility – Repeatable workflows
• Objective metrics – useful for diagnosis, characterization, therapy
  – Augment / Integrate with genomic, proteomic data
• Deterministic process as much as possible
  – Validate
• Easily deployable to most users and easy to maintain
• Scalable
• Human Computer Interface (Human is in the loop)
• Workflow optimization to optimize people time
  – Automate the tedious/boring stuff as much as possible
Challenges of “Ground Truth” / “Gold Standard”

How these are derived/defined matters.
How accurate are the “gold standards”?
Usually derived from human segmentation.
The computer algorithms analyzes images differently from humans.
We may need to collect and process images differently to take advantage of machine analysis.

Collection of raw images specifically for machine learning/analysis will facilitate advancement.
Ground Truth Challenges
Our experience at FNLCR

• Often no “Ground Truth” for the problem at hand
  – Tissue Doppler imaging for Chemotherapy induced Cardio Toxicity
  – Locating and quantifying metastases in mouse models (usually genetically modified)
  – Vessel segmentation for angiography
  – Lymphangiography of HIV/SIV infection
  – Particle picking in Cryo-EM
  – Feature segmentation on digital pathology images
  – Apoptosis quantification via Ultrasound imaging

• Can synthetic data and simulations help address the issue as a refinement step
  – Cryo-EM is a real possibility (Computationally intensive)
  – Generative Adversarial Networks (GAN)
Ground Truth Challenges Facilitating more complete pathology annotation

- There is need for more gold standard annotations
- Currently looking at tools needed for pathologists to label imagery
  - Emory/Kitware HistomicsTK
  - Cytomine
- What about crowdsourcing annotation?

Nuclei classification algorithms (from 3D Slicer) run on uploaded pathology imagery
Research Article

Use of Mechanical Turk as a MapReduce Framework for Macular OCT Segmentation

Aaron Y. Lee,1,2 Cecilia S. Lee,1,2 Pearse A. Keane,2,3,4 and Adnan Tufail2,3,4

completing an average of 5.5 HITs. Each HIT was completed in an average of 4.43 minutes. Conclusions. Amazon Mechanical Turk provides a cost-effective, scalable, high-availability infrastructure for manual segmentation of OCT images.
Use Cases and Observations

Experience using Deep Learning (Segmentation)
Assessing/Understanding Deep Learning Models
Understanding HIV/SIV Infection
DLNN analysis/quantification for RNAscope

- Quantification goals
  - Isolated particles
  - Aggregation
  - Productive Infections
Understanding HIV/SIV Infection
DLNN analysis/quantification for RNAscope

Patch generation for DLNN training, 3184 patches (256x256) in total, un-augmented

Multi-step label generation for DLNN training

cropping

Original

Final Label

DLNN training and prediction

Training Summary:
• Avg. validation score (Dice coef.): ~0.93
• Avg. testing score (Dice coef.): ~0.96

Customized Data Analysis Workflow

Counts, statistics, etc.

Frederick National Laboratory for Cancer Research
DLNN as a component of biomedical research workflows
Understanding HIV/SIV Infection
Deep Learning in Digital Pathology (collagen segmentation)

- Object detection/classification is popular and easier to implement
- Segmentation, on the other hand, is more challenging, demands more accuracy

Image provided by TAC, ACVP, FNLCR

Training data

SVM+postprocessing

DLNN training

Trained Network
U-Net
LCNN (in house)
etc.

Quantification based on DLNN prediction

Original image
Generated label

Color matched images

Collagen index: 253
Number of branches: 3
Length: 24.14
Curvature: 1.12

Collagen index: 192
Number of branches: 131
Length: 528.22
Curvature: 0.60

New images with different stains

Quantification based on DLNN prediction

Color matched images
Training Statistics

- Original images utilized for training
  - Before augmentation: 800 x 800 in size, 275 images
  - After augmentation: 400 x 400 in size, 37,400 images
- Images utilized for Validation
  - 91 images, 800 x 800 in size, divided into four, no augmentation
  - 400 x 400 in size, 364 images
- Images utilized for Testing
  - 91 images, 800 x 800 in size, divided into four, no augmentation
  - 400 x 400 in size, 364 images
- Latest results (single training: 14 hours, 40 epochs, 4x1080 Ti, U-Net variation)
  - Training Accuracy: 96.69% (Dice coefficient)
  - Validation Accuracy: 95.17% (Dice coefficient)
  - Testing Accuracy: 94.65% (Dice coefficient)
DLNN as a component of biomedical research workflows (WSI Tissue Section Collagen Quantification)

**Input Tissue Section Images (Lymph Nodes)**
*Image provided by PHL, LASP, FNLCR*

17040205_0
17040205_1
17040205_2
17040205_3
17040205_4
17040205_5

Whole slide image
*(17040205.svs, 51,791 x 45,864, ~2.3 billion pixels)*

**Preliminary quantification results based on DLNN segmentation correlating to anatomical structures**

ROI 17040205_5, ~0.2 billion pixels

**Number of collagens**

**Avg. len. collagens**

**Avg. num. branches**

Frederick National Laboratory for Cancer Research
Deep Learning Observations

- Training a network is straightforward: DL frameworks are mature and easy to use nowadays
  - But much harder to generate good training data (not so bad for classification/detection)
  - DLNN performance is important, what’s more important is how much knowledge extracted from DLNN
  - Deep Learning can be used even without ‘big data’
  - Deep Learning can be very useful in medical imaging, incremental training data generation is possible solution to solve the scarcity of good labels
    - Incremental means using existing algorithms/ML/manual process to generate small number of labels to train a initial network and use it to predict more labels → labels to be refined to re-train the network to increase number of data points
  - Careful data augmentation on medical images: enhancement should follow the original data distribution so it has to be guided and fit the context
    - Data augmentation not always necessary: in our PDX whole-tumor segmentation task, 76*36 256x256 un-augmented images reached 95% average segmentation accuracy → connect back to ground truth issue
  - Computationally demanding
    - 4 X 1080 Ti GPUs can only train networks with batch size 8 for 400x400 input image size (larger batch size is better)
    - 3D CNNs for volumetric image data demand more resource (roughly one 672x672x40 volume per 64GB V100 card)
Investigating DL Sensitivity

- Computation - IBM Power Systems
  - 2 POWER8 processor modules
    - (8/10 cores, 3.259/2.860 GHz)
  - 4 NVIDIA Tesla P100 GPUs
- 136 Mouse MRIs
  - 130 mice with tumor
  - 6 mice without tumor

- DL Network
  - Convolutional neural network: U-net
  - Learning rate: 3.0e^{-5}
  - Epochs: 80 (2742 images/epoch)
  - Batch size: 8
  - Image size: 400 x 400 pixels (original image size: 672 x 672)
  - No data augmentation
  - Duration per training: 2h 50 min

Each Mouse MRI

- Corresponding label

A single frame from MRI of a mouse

× 36 frames

Corresponding label

× 36 frames

Image provided by SAIP, LASP, FNLCR
Data Selection ("Fixed" data set)

- Six mice without tumor are included in training data only
  - Dice coefficient (evaluation metric) is penalized heavily when mice without tumor are included in validation/testing data
- Fixed data set is utilized for training convolutional neural network (CNN)
  - Repeated the training 10 times
Data Selection (“Random” data sets)

- Six mice without tumor are included in training data only
  - Dice coefficient (evaluation metric) is penalized heavily when mice without tumor are included in validation/testing data
- Data partitioning is repeated 10 times to generate 10 different data sets
  - Repeated the training 10 times
- Per training, the CNN is trained with a different sampling of total data set

Fig. Simple diagrams showing how 10 different data sets are created for training convolutional neural network
Individual Results plotted for Comparison

Normal distribution curves on validation and testing scores

Normal distribution curves for validation scores

- Fixed data
- Randomly selected data

Normal distribution curves for testing scores

- Fixed data
- Randomly selected data
Individual Results plotted for Comparison Initializers
What is the “truth”? How can we tell?

• If we could evaluate images using multiple human evaluators (Expert Evaluation) with different conditions and different machine learning models (Machine Evaluation), then we could compare the distributions and estimate the probability that they represent the same distributions.

• HPC/Advanced Computing could enable this type of evaluation.
Investigating DL Sensitivity (Optimization Effects)

- **Training data**
  - 3184 256x256 RGB training patches
  - B&W training labels created using semi-manual methods
  - No data augmentation

- **Testing data**
  - 120 256x256 manually annotated patches

- **Training Strategy (DLNN tested: U-Net)**
  - Repeat training multiple times (~40) and up to 50 epochs each time
    - In each run, randomly select 75% for training and 25% for validation
    - Other parameters kept the same across multiple runs
  - With vs. without Early Stopping and Reduce Learning Rate on Plateau
    - Learning rate starting at 1e-4
  - Trained networks with <0.9 validation score excluded in reporting
Individual Results plotted for Comparison

Normal distribution curves on testing scores

U-NET Validation vs. Testing (with ES and ReduceLr)

U-NET Validation vs. Testing (without ES and ReduceLr)

Early Stopping and Reduce Learning Rate help to reduce variations in testing scores
Investigating DL Sensitivity (Random Seeds)

- Machine learning algorithms are stochastic in practice
- Fixed seeding
  - Use an arbitrarily selected random seed across all trainings
- Randomized seeding
  - Restart training from the beginning every 10 runs to use new random seeds
Acknowledgements

• Imaging and Visualization Group
  – Hyun Jung
  – Christian Suloway
  – Tianyi Miao
  – Curtis Lisle, KnowledgeVis
  – Xinlian Liu, Hood College

• Our Collaborators at FNLCR, NCI/NIH
Thank you

Questions?
Deep Learning Questions

• How robust / sensitive are the models with respect to training parameters
• Image Augmentation
  – Currently ad-hoc “black box” used to improve training accuracy
  – Research really needed to determine the effect of different augmentation techniques on training accuracy, specificity in biomedical imaging
• Neural Network Architectures
  – Innovation with different connectivity
• Standards
  – Open Neural Network Exchange (onnx) format proposed for support across libraries (https://github.com/onnx/onnx)