Bladder Cancer Diagnosis using Deep Learning

Mauro Damo, Wei Lin, Ronaldo Braga and William Schneider
Agenda

- Who We Are
- Bladder Cancer Diagnosis using Deep Learning Study
- Global Enablement for Cancer Study using Big Data and Deep Learning
Dell Technologies Addresses All Four Transformation Pillars

- Digital Transformation
- IT Transformation
- Workforce Transformation
- Security Transformation
Dell EMC Consulting

Our expert consultants accelerate time-to-value for our customers’ transformations by leveraging our deep knowledge across Dell Technologies

TRANSFORMATIONAL PROGRAM OFFICE

Digital Transformation
- Cloud Native Apps and DevOps
- Application Portfolio Optimization
- Big Data, IoT, Analytics, and Platforms

Workforce Transformation
- VDI and End User Computing
- Digital Workplace Portals
- Communication and Collaboration

IT Transformation
- Multi-cloud Infrastructure and Operating Model
- Data Center Modernization and Migration
- Business Resiliency

P E O P L E  P R O C E S S  T E C H N O L O G Y

#StrataData
Big Data & IoT Consulting Services
Helping customers with their digital transformations

ADVISE
- Data Strategy
- Use Case Prioritization
- Capability Assessment
- Solution Architecture / Design
- IoT Planning
- DaaS Planning and Architecture
- Technology Advisory Services

PLAN
PROGRAM and Organizational Enablement
- Operating Model (Skills, Org, Process)
- Skillset Assessment
- Governance Integration
- Financial Impact Analysis

EXECUTE
- Use Case Development
- Analytics Factory
- Mentorship Program
- Data Science as a Service
- Use Case Operationalization

ANALYTICS from Exploration to Production
- Complex Events Analysis
- Analytical Models Enhancement
- Predictive Analytics
- IoT Analytics
- Machine Learning/Deep Learning
- Text, Audio, Video and Image Analytics
- HPC and GPU Computing

PLATFORM Design, Implementation and Optimization
- Solution Architecture/Design
- PoC, PoV, Tools Validation
- Tech Assessment & Health Check
- ETL/EDW Offload and Migration
- Hadoop Implementation (DAS/NAS)
- Hadoop on Isilon Services
Bladder Cancer Diagnosis using Deep Learning Study
Bladder Cancer is the fourth most common in men

- **430,000 new cases** per year globally

- Estimate of **81,191 new bladder cancers in US** and **fourth most common** in men

- Direct medical **cost** of bladder cancer care was **$125B** in 2010 globally

- Cost of muscle-invasive bladder cancer is **$150k** and early stage cancer (first 2 yrs) is **$10k per patient** globally


Our Proposal

- Multinomial classification of primary tumor that can recognize bladder cancer in Magnetic Resonance Images without human intervention

- TMN classification
  - **Tumor**: How large is the primary tumor? Where is it located?
  - **Node**: Has the tumor spread to the lymph nodes?
    - If so, where and how many?
  - **Metastasis**: Has the tumor spread to the lymph nodes?
    - If so, where and how many?

- **Focus on primary tumor**

- Tracked 4 different types of primary tumors of bladder cancer: **T2a, T2b, T3a and T4a**
Hardware and Software Stack

Intel Xeon E5-2680 @ 2.7GHz with 8 cores and 384 GB

NVIDIA GRID K2 with 2 GPUs GK104 with 1.536 cores per GPU and 4 GB per gpu RAM

Script Language with the following packages. Main data transformation
- Python
- Numpy
- Matplotlib
- OS
- Jupyter

Deep Learning package:
- Neural Network Build and Design
- Mini batch process
- Functions and Loss Functions
- Algebra Computation
- Optimization Process

Image transformation package:
- Shrink
- Binarization

Medical Image file standard
Library to access the Metadata and Data (Pixels)
How MRI device scans patient organs

- MRI device can scan the patient in 3 gradients - each of them create an image with a different perspective
  - Coronal
  - Transverse
  - Sagittal

Data Set and Images

- 5,019 Magnetic Resonance Images of the pelvic region from patients
- No previous image selection for all the images in the session
- No image cropping or regional detection was done in raw data
- All patients had cancer - our goal was to detect the class of the tumor in different patients

Figure 6 - Extent of primary bladder cancer

Primary Tumor (T)

- T<sub>x</sub>: Primary tumor cannot be assessed
- T<sub>0</sub>: No evidence of primary tumor
- T<sub>a</sub>: Non-invasive papillary carcinoma
- T<sub>is</sub>: Carcinoma in situ: "flat tumor"
- T<sub>1</sub>: Tumor invades subepithelial connective tissue
- T<sub>2</sub>: Tumor invades muscle
- T<sub>2a</sub>: Tumor invades superficial muscle (inner half)
- T<sub>2b</sub>: Tumor invades deep muscle (outer half)
- T<sub>3</sub>: Tumor invades perivesical tissue
- T<sub>3a</sub>: Microscopically
- T<sub>3b</sub>: Macroscopically (extravesical mass)
- T<sub>4</sub>: Tumor invades any of the following: prostate, uterus, vagina, pelvic wall, abdominal wall
- T<sub>4a</sub>: Tumor invades prostate, uterus, vagina
- T<sub>4b</sub>: Tumor invades pelvic wall, abdominal wall
Data Transformation and Tensors

MRI IMAGES
Gray Scale
512x512

Shrink subsample

MRI IMAGES
Gray Scale
256x256

Threshold Pixels

Transformed MRI IMAGES
Binarization
256x256

Transform in tensor with (256,256,1) for a predictors and a tensor of (1,4,1) of labels

Image transformation was needed to fit in computation power and increase accuracy
Neural Network Architecture

- Used a 6 layer convolution neural network
- 4 layers of 2d convolution strides = \([1,1,1,1]\) and padding='sample'
- 2 layers of Full Connected with dropout
- Softmax layer for multinomial classification
- Max Pool with 2x2
- Relu as activation function
Results

- Classification outcomes are related to 4 classes: T2a, T2b, T3a and T4a

- Using the ConNet, Top 1 accuracy increases achieving 81.30%

- Baseline using a Multinomial Logistic Regression we achieved Top 1 accuracy 72.27%
Lessons Learned

- GPU and CPU memory are more relevant in your hardware than cycles
  - OOM errors are very common when we use medical data; unless your model take weeks to run, it is better to have more memory to fit all your weights initialization and mini batch process
  - If your model takes weeks to run, it is better to improve memory and cycles or use a distributed platform

- Code Design
  - Image processing was a very time demanding phase: after applying several different types of image filters, we needed to train the CNN and test the model to see the Top 1 accuracy of the model which took time
  - Convolution Neural Network: we tried different CNN approaches, including 2,3,4 and 5 layers. How large is your CNN, it seems that it gives you better results, similar to ResNet.
Conclusion

- Convolution Neural Network Architecture has a positive path in Medical Images Diagnosis - an increase of accuracy from 72.3% to 81.3% shows potential to explore

- Some techniques to generalize CNN in Data Science:
  - Availability of more data in early stages cancer and record of clinical checkpoint with patients
  - Application of R-Fast-CNN with other CNN architecture (like ResNet) for image segmentation and classification

- Some techniques to generalize in medical domain:
  - Cover more types of primary tumors
  - Train independently CNN models for all planes (coronal, transverse and sagittal) using a voting criteria
Global Enablement For Cancer Study using Big Data and Deep Learning
Samples of ML/DL Algorithms and Stack Selection

Criteria
- Identify inputs characteristics (continuous variables, categorical variables, text, log, image, video, voice)
- Verify Target variable (continuous vs. categorical and available vs. not available)
- Decide approach (supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning)

Rules of Thumbs:
- Machine Learning Stack, Libraries
- Operating Big Data Ecosystem and Development
- Speed, Visualizations, Production
Dell EMC Medical Analytics Factory Model 2.0

**FACTORY LEAD**

**BUSINESS/TECH ANALYSTS**

**ANALYTICS PIPELINE**

**TESTERS**

**DEPLOYMENT LEAD**

### Prioritize
- Business owner communication
- Prioritize tasks
- Workforce and environment planning and provision
- Capacity provision
- Track backlog
- Progress communication

### Classify
- Assess Complexity
- Functional evaluation to existing modules.
- Consolidate Determination
- Design review
- Common utility
- Consolidated use cases
- Workforce provision and review

### Perform
- Descriptive Analytics (D): qualify use case data, extract data quality, density, distributions, skewness, and variable correlations
- Exploration Analytics (E): Explore hypotheses related to use case. Generate relationships among hypotheses, verify hypotheses and generate measurements.
- Predictive Analytics (P): perform root cause analysis, events linkages, outliers detection, forecast measurements by ML or DL.
- Prescriptive analytics (P): Convert Analytics output into business actionable, Plan A/B Tests, and tracking outcomes.

### Test
- Conduct
  - Analytic model code test
  - API wrapper code test
  - Performance test
  - Unit Test
  - User Test, Business APP Integration test
  - "Execute A/B Business Test"

### Deploy
- Coordinate and Manage
  - Rollout planning
  - Training
  - Documents
  - Deployment planning, business app

### PROCESSES MINING

Statistics of DevOps KPIs: Counts/ time based distributions of Pause, Queue, Checkout, Build Automation, Unit Test, Code analysis, Build management, Security Test, Deployment, release frequency, commit frequency, Deployment frequency, Change failure rate, mean lead time to change, mean time to recover, ...

Processes Analytics: Root Cause, Cauacity Analysis, Data Center predictive maintenance (Hardware, software, APPs), Customer ROI trend, .. Complex event process improvement, ..

Reasons, Priority, Backlog, funding, Analytics Factory KPIs/Metrics (AKM)

Duration, impacts, expected performance, common design patterns, .. AKM

Data sources quality, sys owner, ready capacity, accept/reject AKM

Hypothesis trending, data enriching AKM

Accept/reject, ROC tracking, re-train performance.. AKM

Business process integration counts .. AKM

Test tracking, tickets trending, performance distributions, .. AKM

Deployment, user experience, Expected vs. Real ROI tracking, ..AKM

#StrataData
Process Mining Example

1. Increase trend of change failure rate induces
   - Low production commit frequency
   - Low deployment frequency
   - Low mean lead time to changes
Step 1: Establish hypothesis

Hypothesis
CMs have “quality production ranges” where production outside of those ranges have **Lower** than acceptable quality problems

Step 2: Identify and quantify predictive variables

<table>
<thead>
<tr>
<th>Variables</th>
<th>F(x)</th>
<th>Var σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>CM Recent Performance Score</td>
<td>.67</td>
<td>2.50</td>
</tr>
<tr>
<td>Supplier x Component Quality Score</td>
<td>.62</td>
<td>3.25</td>
</tr>
<tr>
<td>Product Complexity Score</td>
<td>.59</td>
<td>1.25</td>
</tr>
<tr>
<td>Test Suite Score</td>
<td>.58</td>
<td>3.60</td>
</tr>
<tr>
<td>Seasonality</td>
<td>.51</td>
<td>1.14</td>
</tr>
<tr>
<td>Platform Stability Index</td>
<td>.49</td>
<td>4.55</td>
</tr>
<tr>
<td>Extreme Weather</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Step 3: Build CM Profiles

Step 4: Monitor data feeds against profiles to flag anomalies
- BU develop x Software modules Component DPPM
- Supplier x Component TST
- PCN
- ...
- Day of week
- Local weather
- CM Newsfeeds
- Local economy

Step 5: Refine Profile variables
- Reduce variance
- Add new variables
- Delete variables

Step 6: Publish CM Quality and Rework Scores & Recommendations

#StrataData
Analytics Factory Efficiency Logical N-Tiers View

LOGICAL VIEW

WORKSPACE
DATA CURATOR
CATALOG
INGEST

WORKSPACE
DATA SCIENTIST

USERS
BUSINESS ANALYST
IT ADMINISTRATOR

APPLICATION DEVELOPMENT
ANALYTICS FABRIC
DATA LAKE
VIRTUALIZATION
NETWORKING
COMPUTE
HPC
STORAGE

POLICY-DRIVEN DATA LAKE MANAGEMENT
ENVIRONMENT
ANALYTICS
DATA SETS

DATA GOVERNOR
POLICIES
LINEAGE
QUALITY
SECURITY

#StrataData
Dell EMC Medical Analytics Factory with Dell EMC WWH* for Global Deployment and GDPR Compliance

- Share “Knowledge” without disclosing PII and PHI
- Operating Big Data ecosystem and development facilitated by processing management
- Distributed algorithm design and testing (step 1, 2) and analytics model training and sharing deployment

Q&A

Wei Lin – w.lin@dell.com       Mauro Damo – mauro.damo@dell.com
This slide represents Dell Technologies’ operating structure. Our financial reporting structure consists of three business units: CSG, ISG, and VMware. Our other businesses include the results of RSA, Pivotal, Secureworks, and Boomi.