Big Image-Omics Data Analytics for Clinical Outcome Prediction

Junzhou Huang, Ph.D.
Associate Professor
Dept. Computer Science & Engineering
University of Texas at Arlington
Morphology and Prognosis

- **Integration:**
  - Connections between morphology and prognosis
  - **How:** integrate pathological image data and molecular profiling data to learn this connection?
Clinical Outcome Prediction from Heterogeneous Cancer Data

- **Problem:**
  - Subtype Recognition
  - Survival Prediction

- **Data:**
  - Pathological Image
  - Gene Mutation
  - CNV
  - mRNA Expression
  - Protein Expression

- **Cohort:**
  - TCGA (The Cancer Genome Atlas)
  - NLST (The National Lung Screening Trial)
  - UT lung SPORE cohort.
Pipeline Overview

Subtype Cell Detection

• Motivation:
  – Different cell types (tumor cells, stromal cells, lymphocytes) play different roles in tumor growth and metastasis
  – Accurately classifying cell types is a critical step to better characterization of tumor growth and outcome predictions.

• Traditional Cell Detection Methods[1]:
  – Pros: easily implemented and interpreted; faster
  – Cons: performance is not good enough

• Deep Learning Cell Detection methods[2]:
  – Pros: better detection performance.
  – Cons: Slow;

Deep Learning for Subtype Cell Detection

- We designed a special structure for subtype cell detection:
  - **Shared Convolution Weights**: the cell/non-cell deep convolution neural network and subtype deep convolution neural network share all convolution weights to avoid the insufficiency and imbalance of the subtype cell patches.
  - **Sparse Kernel**: introducing the d-regularly sparse kernels to elimination all the redundant computation and to speed up the detection process.

Results on Subtype Cell Detection

• Detection Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 score</th>
<th>Times(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NERS[1]</td>
<td>0.7990</td>
<td>0.6109</td>
<td>0.6757</td>
<td>31.47</td>
</tr>
<tr>
<td>RLCCD[2]</td>
<td>0.7280</td>
<td>0.8030</td>
<td>0.7759</td>
<td>52.89</td>
</tr>
<tr>
<td>Proposed</td>
<td>0.8029</td>
<td>0.8683</td>
<td>0.8215</td>
<td>0.7147</td>
</tr>
</tbody>
</table>

• Subtype Detection Results
  – Subtype Classification Neural Network Accuracy: 88.64%
  – Accuracy of Detected Cells: 87.18%
    – Lymphocytes Accuracy: 88.05%
    – Stromal Cell Accuracy: 81.08%
    – Tumor Cell Accuracy: 87.39%

Our method has better performance in terms of both accuracy and computational time

Is lung cancer subtype cells detection easy?

The size of one sample image is $13483 \times 17943$ (usually larger), while traditional cell detection methods are still dealing with images with size $\sim 500 \times 500$.

Cell density could be very high!

Image size: $512 \times 512$
Pixel scale: $0.25\mu m/pixel$
Acceleration (1): Sparse Kernel

Sparse Kernel is used to eliminate all redundant computations in convolutions.

The yellow area will be calculated several times since it will appear in several overlapping patches.

We take the whole image as input and reuse the convolution operation result in the detection, which could roughly accelerate several hundred times depending on the sliding window size.
Acceleration (2): Prefetching Technique

We use Asynchronous Prefetching Technique to
Acceleration (3): Cluster Computing

- **Single-node Multi-GPU Computing**
  - Communication Through PCI-e bus
- **Multi-node Computing**
  - Communication Through Network

- The data is mapped to a high-performance large-scale Network File System
- Only the coordinates are communicated in the distributed system, which makes our framework **scalable and communication-efficient in the cluster computing system.**
Results on the Single Machine

Time Comparison in Different Whole-slide Images

<table>
<thead>
<tr>
<th>Image Name (Dimension)</th>
<th>1 GPU</th>
<th>2 GPUs</th>
<th>3 GPUs</th>
<th>4 GPUs</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLSI0000105 (13483 × 17943)</td>
<td>71.43</td>
<td>38.81</td>
<td>26.89</td>
<td>20.88</td>
</tr>
<tr>
<td>NLSI0000081 (34987 × 37879)</td>
<td>366.74</td>
<td>194.99</td>
<td>131.30</td>
<td>99.20</td>
</tr>
<tr>
<td>TCGA-05-4405 (83712 × 50432)</td>
<td>1502.16</td>
<td>800.24</td>
<td>529.00</td>
<td>449.94</td>
</tr>
<tr>
<td>TCGA-35-3615 (62615 × 133335)</td>
<td>2953.99</td>
<td>1519.57</td>
<td>1100.32</td>
<td>861.17</td>
</tr>
<tr>
<td>TCGA-38-4627 (65033 × 149642)</td>
<td>3385.28</td>
<td>1773.11</td>
<td>1216.80</td>
<td>972.36</td>
</tr>
</tbody>
</table>

Test Machine:
- CPU: Intel(R) Core(TM) i7-5930K CPU @ 3.50GHz
- RAM: 64 Gigabytes
- GPU: 4 Nvidia Titan X GPUs
- HDD: Samsung 950 Pro Solid-State Drive

We’re able to detect cells in a **13483 × 17943** image within only **20** seconds, on a **single machine**! (4000 times acceleration! Larger, more acceleration!)
More Results on the GPU Clusters

Time Comparison on TACC Stampede Cluster

<table>
<thead>
<tr>
<th>Image Name (Dimension)</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>16</th>
<th>32</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLSI0000081 (34987 × 37879)</td>
<td>520.94</td>
<td>266.06</td>
<td>143.99</td>
<td>77.16</td>
<td>44.10</td>
<td>26.03</td>
</tr>
<tr>
<td>TCGA-05-4405 (83712 × 50432)</td>
<td>1820.08</td>
<td>945.77</td>
<td>508.23</td>
<td>271.02</td>
<td>155.39</td>
<td>86.31</td>
</tr>
<tr>
<td>TCGA-35-3615 (62615 × 133335)</td>
<td>3558.48</td>
<td>1834.00</td>
<td>944.91</td>
<td>487.47</td>
<td>266.35</td>
<td>147.07</td>
</tr>
<tr>
<td>TCGA-38-4627 (65033 × 149642)</td>
<td>4151.56</td>
<td>2107.46</td>
<td>1086.53</td>
<td>559.28</td>
<td>293.98</td>
<td>155.87</td>
</tr>
</tbody>
</table>

With **32 Nvidia Tesla K20 GPU nodes**, the benchmark of our framework in a $65033 \times 149642 (10^{10})$-pixel image is only **155 seconds**.

(≈10,000 times acceleration!)

TACC Stampede Cluster:
https://www.tacc.utexas.edu/stampede/
Pipeline Overview
Biomarker Discovery for Survival Prediction

- **Data set**
  - The National Lung Screening Trial (NLST): 144 ADC, 113 SCC

- **Predictive models**
  - Multivariate Cox proportional hazards model with Lasso
  - Component-wise likelihood based boosting (CoxBoost)
  - Random survival forest (RSF)

- **Experiment Set**
  - Compare with the state-of-the-arts framework in lung cancer
  - Compare performances on different types of data (CNV, mRNA, microRNA and Protein expression) on TCGA-LUSC


Results

- Survival Prediction

A significant difference can be seen (smaller p-value) in the proposed framework

A significant difference can be seen (smaller p-value) in the proposed framework.

Results

- Random Experiments (50 splits)

Concordance index (C-index):

1 indicates perfect prediction accuracy, 0.5 is as good as a random guess.

Integration with Molecular Data
Existed Methods

• **Supervised: Univariate, Forward stepwise selection, Lasso, etc.**
  – **Pros:** outcome prediction related features could be potentially selected
  – **Cons:** the correlations among the bi-modal data features are not available

• **Unsupervised: Conditional Gaussian graphical models**
  – **Pros:** the mappings between bi-modal data are learned
  – **Cons:** noisy output with respect to the clinical outcome prediction due to the lack of supervised information

Supervised CGGM (SuperCGGM)

Pathological Image

Genetic Expression Signatures

Clinical Outcome

Feature Extraction

Expression Values

Expression Values

0 2.3

0.53 ⋯ 0.02

⋮ ⋱ ⋮

0.01 ⋯ 0.39

0.90 ⋯ 0.16

⋮ ⋱ ⋮

0.41 ⋯ 0.05

Supervised

Y

Supervised

T

Survival Time

X

P(Y|X)

Clinical Outcome

Pathological Image

Genetic Expression Signatures

\[
\begin{bmatrix}
0 & 2.3 \\
\vdots & \vdots \\
1 & 0.5
\end{bmatrix}
\]

e: death (1) or live (0)

t: observation time
Experiment

• Data Set:
  – 111 ADC lung cancer patients from UT Lung SPORE,
  – UT Lung SPORE: a plan to identify and understand the molecular "hallmarks of lung cancer" and then translate this information into the clinic for early detection, prevention, prognosis, and the selection and/or development of new treatments for lung cancer.
  – Genetic Expression Values: 21905 dimensions
  – Image Features: 1794 dimensions
  – 2/3 training, 1/3 testing
Survival Prediction

- Survival prediction performances of Univariate, Lasso, Ridge, FSS, PCA, SuperPCA, PLS, SPACE, CGGM and SuperCGGM (smaller p and larger CI are better)

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Image p</th>
<th>Image CI</th>
<th>Gene p</th>
<th>Gene CI</th>
<th>Integration p</th>
<th>Integration CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Univariate</td>
<td>0.42</td>
<td>0.532</td>
<td>0.19</td>
<td>0.553</td>
<td>0.19</td>
<td>0.562</td>
</tr>
<tr>
<td>Lasso</td>
<td>0.06</td>
<td>0.595</td>
<td>0.12</td>
<td>0.564</td>
<td>0.08</td>
<td>0.592</td>
</tr>
<tr>
<td>Ridge</td>
<td>0.08</td>
<td>0.592</td>
<td>0.07</td>
<td>0.576</td>
<td>0.07</td>
<td>0.616</td>
</tr>
<tr>
<td>FSS</td>
<td>0.12</td>
<td>0.604</td>
<td>0.10</td>
<td>0.540</td>
<td>0.14</td>
<td>0.593</td>
</tr>
<tr>
<td>PCA</td>
<td>0.11</td>
<td>0.607</td>
<td>0.10</td>
<td>0.536</td>
<td>0.15</td>
<td>0.587</td>
</tr>
<tr>
<td>SuperPCA</td>
<td>0.04</td>
<td>0.651</td>
<td>0.05</td>
<td>0.595</td>
<td>0.05</td>
<td>0.606</td>
</tr>
<tr>
<td>PLS</td>
<td>0.08</td>
<td>0.634</td>
<td>0.11</td>
<td>0.541</td>
<td>0.09</td>
<td>0.622</td>
</tr>
<tr>
<td>SPACE</td>
<td>0.08</td>
<td>0.622</td>
<td>0.06</td>
<td>0.566</td>
<td>0.08</td>
<td>0.622</td>
</tr>
<tr>
<td>CGGM</td>
<td>0.07</td>
<td>0.640</td>
<td>0.08</td>
<td>0.607</td>
<td>0.03</td>
<td>0.630</td>
</tr>
<tr>
<td>Proposed SuperCGGM</td>
<td><strong>0.03</strong></td>
<td><strong>0.660</strong></td>
<td><strong>0.05</strong></td>
<td><strong>0.613</strong></td>
<td><strong>0.009</strong></td>
<td><strong>0.691</strong></td>
</tr>
</tbody>
</table>
Summary and Future Work

**Summary**
- Big imaging-genomic data analytics framework for lung cancer clinical outcome prediction
- Scalable deep learning algorithms for big image data analytics with 10,000 times acceleration
- Novel integration method for survival prediction from image-gene data

**Future Work**
- Further improve imaging analysis (segmentation, feature extraction, etc)
- Closely work with lung cancer pathologists to develop better algorithms and more clinically meaningful features, which are guided by clinical knowledge
- Predict drug response using pre-clinical and clinical samples.
Thank You!