Parallel Training of Large Knowledge Graph Convolutional Networks

Hong-Jun Yoon, PhD

2020 Performance, Portability, and Productivity in HPC Forum

September 2020

ORNL is managed by UT-Battelle, LLC for the US Department of Energy
Joint Design of Advanced Computing Solutions for Cancer (JDACCS4C)

- US Department of Energy/National Cancer Institute Collaboration

- Three Pilots
  - Pilot 1: Cellular
  - Pilot 2: Molecular
  - Pilot 3: Population-level (ORNL leading)

- Our Team
  - NLP
  - Information Extraction
  - Knowledge Discovery
  - Hypothesis Testing
Data Sources

• NCI Surveillance, Epidemiology, and End Results (SEER) Program
  – Since 1973
  – 450,000+ cases / year
  – 1/3 US population

*Subcontract under New Mexico
**Three regions represent the state of California: Greater Bay, Los Angeles, and Greater California
Cancer Pathology Reports

- Automated information extraction
  - Replace manual or rule-based approaches
  - Scalable training of solutions
  - Deploy API to SEER registries
Under-represented Classes

- Rare cancers
  - Low incidence
  - Low number of training samples
  - Not enough to train our DL models
  - Low classification accuracy

- Solution
  - Import external knowledge sources
  - Knowledge graph, graph convolutional networks
UMLS Concept Relation Graph

Concepts 80,047,154

# of edges 4,177,638

Nodes (CUI's)

- CHD: Child Relationship 6,287,296
- PAR: Parent Relationship 6,287,296
- RB: Broader relationship 1,907,427
- QB: Can be qualified by 609,244
- RN: Narrow relationship 1,907,427
- RO: Other 21,578,246
- SY: Asserted synonymy 6,015,510
- SIB: Sibling relationship 32,303,374
- AQ: Allowed Qualifier 609,244
- RQ: Possibly synonym 2,524,090

96 Data Sources
- MSH - 3,082,856
- ICD9CM - 219,834
- ICD10PCS – 1,263,764
- SNOMEDCT – 6,599,872
- CCS – 57,718
- CSP – 179,448
- GO – 2,498,396
- OMIM – 618,308
- LNC – 3,717,352
- MDR – 2,022,824
...
"This is an 81-year-old female with a history of metastatic melanoma with…"

"Status post excision of skin cancer on left arm."
Graph Convolution with Large Knowledge Graph(s)

• Loosely-coupled: Cluster GCN
  – Divide big one into multiple small dense graphs
  – Concatenate decisions from the multiple GCNs

• Tightly-coupled: Model-parallel GCN
  – Divide one big adjacency matrix
  – Communication overhead
Work In-Progress

- Medical document classification using CUIs
  - Disambiguation of terms
  - Abstraction of various expressions

- GCN
  - Matrix multiplication – GPU-friendly
  - Competitive/higher task performance

- Knowledge Graph
  - Big adjacency matrix – too big to fit one GPU
  - Two approaches
    - Cluster GCN
    - Model-parallel, distributed GCN
Thank you!

- Questions?