ML Systems

Saurav Sengupta, Zachary Jacoakes

Image Processing

- Current neuroimaging processing pipelines utilize parallelization
- Raw signal to usable image transformation is resource intensive
  - K-space to brain-space
  - Per-subject basis
  - One node per subject scan; often many scans
- Entire compute clusters dedicated to neuroimaging research

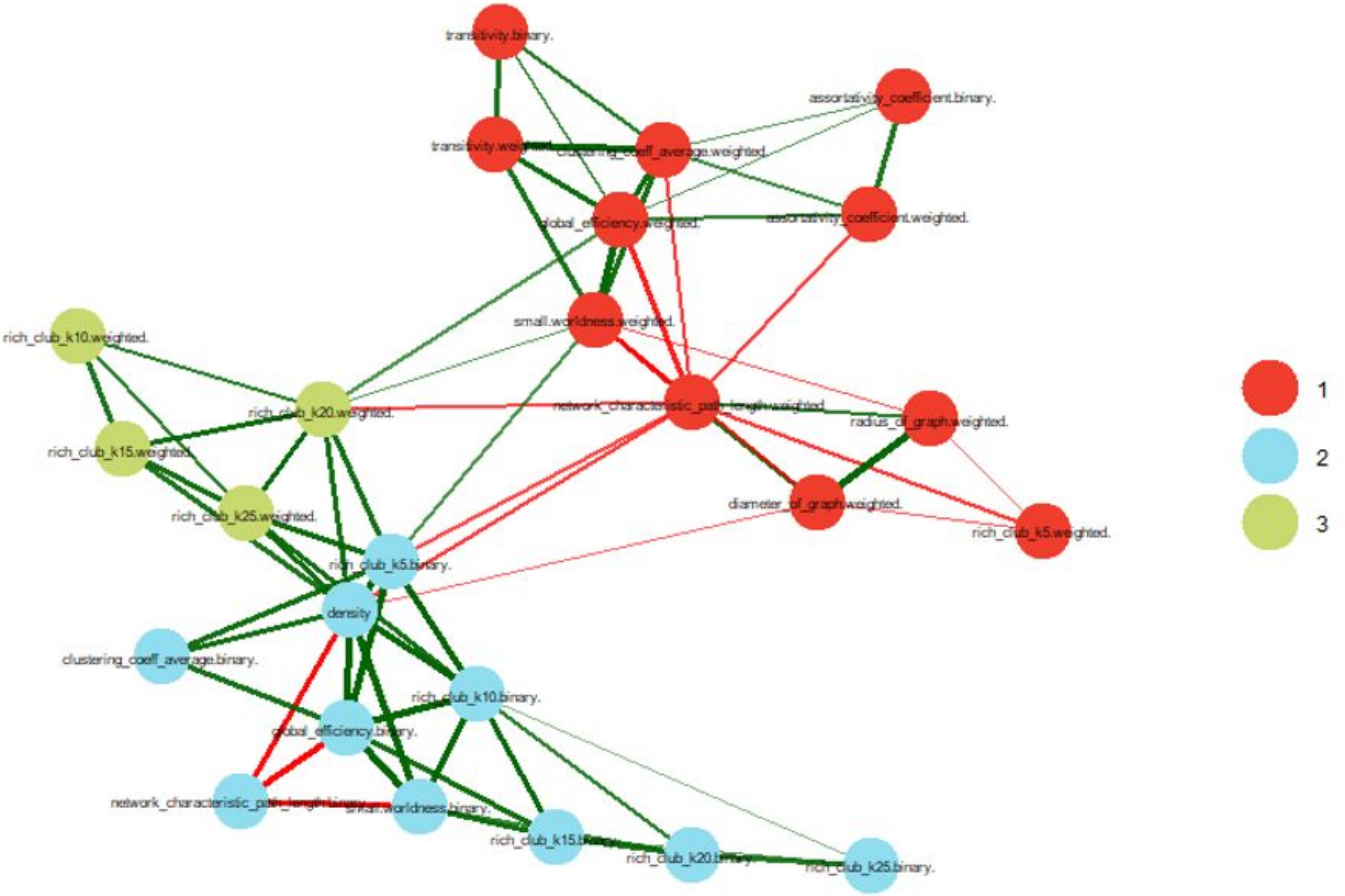
Secondary Post-Processing

- Results in hundreds of features from a single scan
  - Network efficiency, isolated brain region metrics, functional connectivity
- Usually way fewer subjects than features (p >> n)
- Compute clusters are not often leveraged for feature engineering
  - Most dimensionality reduction uses ‘region of interest’-type narrowing
  - Eliminating the signal in the brain that depends on other regions
- Proposal: data-centric feature engineering on public, large-scale neuroimaging projects
  - Human Connectome Project (n > 2000), UK Biobank (n > 50,000), ABCD (n > 10,000)

HUMAN Connectome PROJECT

biobank uk
Enabling scientific discoveries that improve human health

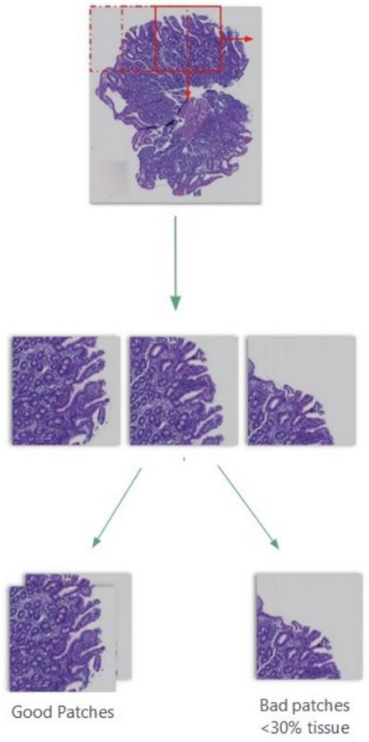
Adolescent Brain Cognitive Development®
Teen Brains. Today's Science. Brighter Future.



transitivity.binary.
assortativity_coefficient.binary.
transitivity.weighted.
clustering_coeff_average.weighted.
global_efficiency.weighted.
assortativity_coefficient.weighted.
small.worldness.weighted.
rich_club_k10.weighted.
rich_club_k20.weighted.
network_characteristic_path_length.weighted.
radius_of_graph.weighted.
rich_club_k15.weighted.
diameter_of_graph.weighted.
rich_club_k25.weighted.
rich_club_k5.weighted.
rich_club_k5.binary.
density
clustering_coeff_average.binary.
rich_club_k10.binary.
global_efficiency.binary.
network_characteristic_path_length.binary.
small.worldness.binary.
rich_club_k15.binary.
rich_club_k20.binary.
rich_club_k25.binary.
1
2
3

Problem

- Medical images can be very high resolution (50,000 X 50,000 pixels on average).
- Modern GPUs cannot fit images of this size in their memory.
- Have to break up images into patches for training - we lose context of the larger image.



Proposed solutions and project ideas

- Distributed data parallelism: Can we develop a method to efficiently distribute the data across multiple GPUs
- Use data compression techniques to compress the image while preserving context of the larger image.
- Use mixed precision training and look at memory consumption.