ARXIV

Automated Subject Indexing

Eliot Kmiec

Cornell University Library

arXiv

Eliot Kmiec

Data Scientist

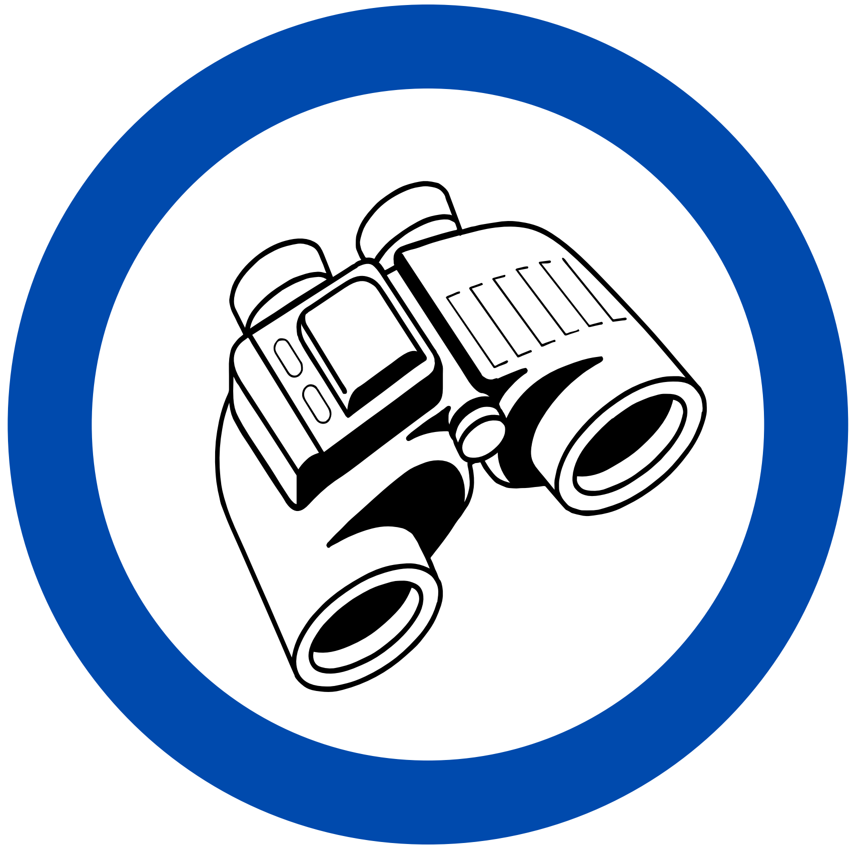
About Me

- Augustana College 2020
  - Biochemistry, Public Health
- Fields of Interest:
  - Bioinformatics
  - Cloud Computing

Why Automate Database Indexing?

- Subject headings are powerful indexing tools
- Labor intensive
- Ongoing work to automate MeSH for PubMed
- Saves Money, Streamlines Research
- ***Bring this technology to arXiv***

Implementing Automated Indexing



Supervised Arm

Identifying the subject of an article

Unsupervised Arm

Identifying new subjects for the database

Approach to Modeling

NLP with RNNs

Data Structure

- Abstracts : Subject Headings
- Available from Kaggle
- 1.5 mil abstracts

Technical Hurdles

- Data Sparsity/High Dimensionality
- Numerous Categories/Subjects

Model Architecture

- Recurrent Neural Networks
- Retain more info VS. traditional models

Using RNNs for NLP

Training Hurdles

The Good

- RNNs scale well
- Accurate even with few examples

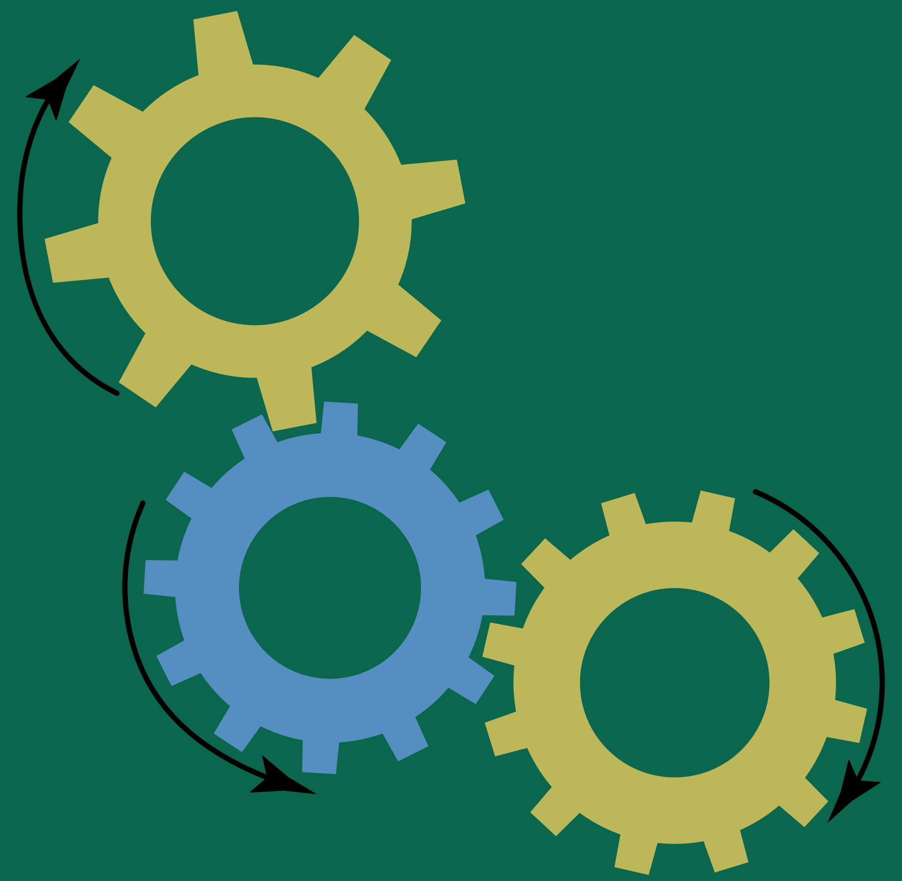
The Bad

- Lots of parameters
- Lots of resources to train

The Ugly

- Prone to gradient instability
- Initial models had "exploding gradient"

The Modeling Process



Gated Recurrent Units

- Basis of the model
- Hidden State
- Gate controlled

Gradient Explosion

- Gradient Clipping
- Batch Normalization
- Categorical Hinge Loss
- AMSgrad optimizer

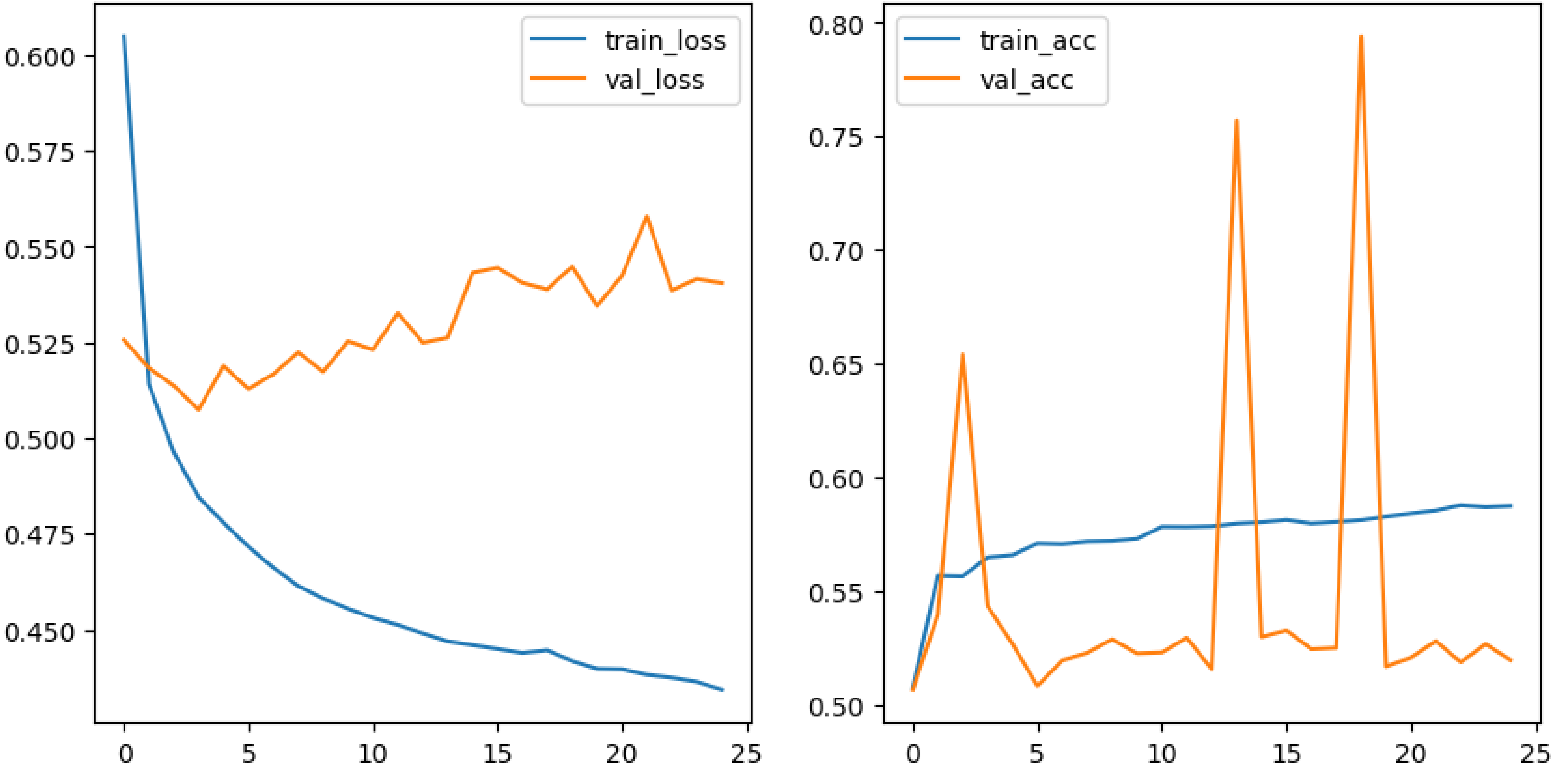
Big Data and Embedding

- Abstracts --> Matrices
- Initially over 100 million params
- (est. 1 month training)

Resource Constraints

- AWS EC2
- GPU access denied
- **Used a subset of the 10 most common subjects**

Model Training & Gradient Instability



Loss

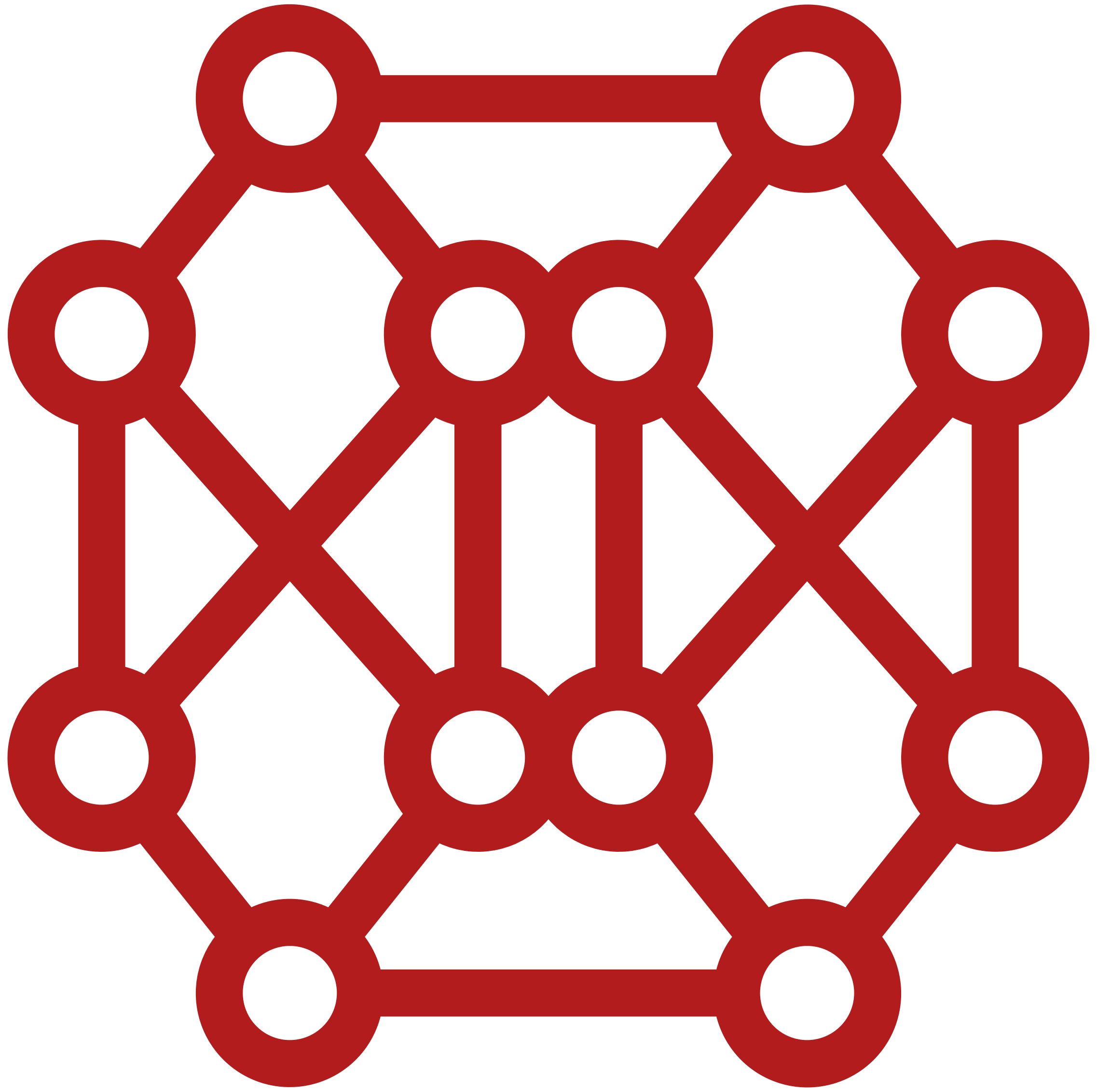
Accuracy

Epochs

Accuracy	*Loss*	*AUC Score*
51.7%	0.541	0.395

Test Scores and Remaining Issues

- 10-classes with 18% dummy model accuracy
- Varying degrees of convergence for most models
- MeSH studies have shown near 80% accuracy



Next Steps

- Parallelize training on GPU cores
- Further work on stabilizing the gradient
- Expand model training to all subjects

Questions?

Project Github

https://github.com/ek775

Eliot Kmiec