Anik Saha

Graduate Student | Machine Learning | NLP
518.428.2758 (cell) • sahaa@rpi.edu • aniksh.github.io

Synopsis

Graduate student with academic and industrial research experience in *Machine Learning* and *Natural Langugae Processing* leading to first-author *publications*. Key skills and experiences:

- Research at **IBM** on **causality extraction** and **document understanding**
- Training and fine-tuning **large language models** like BERT and GPT with **PyTorch**
- **Disributed training** over multiple nodes in RPI-IBM **supercomputers**
- Implementing methods for **domain adaptation** and **knowledge distillation** in low resource scenario

Education

- **M.S. Electrical Engineering** **Aug 2023**
  Rensselaer Polytechnic Institute; GPA: 3.79/4 *Troy, NY*
- **B.S. Electrical and Electronic Engineering** **Sep 2015**
  Bangladesh University of Engineering and Technology; GPA: 3.90/4 *Dhaka, Bangladesh*

Relevant Experiences

- **IBM Research** **Yorktown Heights, NY**
  Summer Research Extern *May 2022 - Aug 2022*
  - Improved domain adaptation performance of neural models for causality extraction from text.
  - Implemented span-based and sequence-tagging models with Pytorch and Huggingface library.
  - Evaluated the effect of pre-training with masked language modeling task on domain adaptation.
  - Introduced task specific output measures in the adversarial domain adaptation method.

  Summer Research Extern *May 2021 - Aug 2021*
  - Incorporated linguistic information in the transformer architecture for causal relation extraction.
  - Converted dependency parse relationships to attention mask for transformers like BERT.
  - Integrated constituency parse information in transformer models to improve span detection.

  Summer Research Intern *Jun 2020 - Aug 2020*
  - Developed transformer models for multimodal information extraction from business documents.
  - Trained LayoutLM on scanned documents to learn textual and 2D positional embeddings.
  - Improved model performance by fine-tuning the trained model to predict the 2D coordinates.
- **Rensselaer Polytechnic Institute** **Troy, NY**
  Research Assistant *Jan 2019 - Aug 2023*
  - Improved word sense induction performance of multi sense embeddings.
  - Developed a multi-stage knowledge distillation method from contextual BERT embeddings to word sense embeddings.
  - Collaborated with IBM Research in the Cognitive and Immersive Systems Laboratory on document retrieval in neural embedding space using simple siamese networks.
  - Adapted sequence tagging and span based models for the causal information extraction task.
  - Evaluated causality extraction performance of neural models on data sets from different domains.
  - Implemented domain adaptation methods for pre-trained transformer models to extract causal information from text.

Teaching Assistant *Aug 2017 - Dec 2018*

- Held office hours, developed assignment solutions and graded assignments for the Introduction to Machine Learning course.

- **Semion Inc.** **Dhaka, Bangladesh**
  Machine Learning Researcher *Sep 2016 - Jul 2017*
  - Developed deep learning models for sentiment analysis of large documents.
  - Utilized distributed computing techniques to speed up training.

- **Daffodil International University** **Dhaka, Bangladesh**
  Lecturer, Department of Electrical and Electronic Engineering *May 2016 - Aug 2016*
  - Taught Introductory Computer Programming, Analog Electronics and Electric Machines.

Skills

Programming Languages: Python, Bash scripts, MATLAB

Version Control: Git

Deep Learning Frameworks: PyTorch, Tensorflow

Machine Learning Tools: Numpy, Scipy, Scikit-learn, Pandas

NLP Tools: NLTK, CoreNLP, Spacy, Gensim

Publications

Anik Saha, Alex Gittens, Jian Ni, Oktie Hassanzadeh, Bulent Yener, and Kavitha Srinivas. Spock@ causal news corpus 2022: Cause-effect-signal span detection using span-based and sequence tagging models. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 133–137, 2022a. URL `https://aclanthology.org/2022.case-1.18/`.

Anik Saha, Jian Ni, Oktie Hassanzadeh, Alex Gittens, Kavitha Srinivas, and Bulent Yener. Spock at fincausal 2022: Causal information extraction using span-based and sequence tagging models. In *Proceedings of the 4th Financial Narrative Processing Workshop@ LREC2022*, pages 108–111, 2022b. URL `https://aclanthology.org/2022.fnp-1.17/`.

Anik Saha, Catherine Finegan-Dollak, and Ashish Verma. Position masking for improved layout-aware document understanding. In *Document Intelligence Workshop at KDD*, 2021. URL `https://arxiv.org/abs/2109.00442`.

Academic Projects

- *Neural Abstractive Summarization with Attention Mechanism* *Spring 2019*
  - Evaluated the pointer-generator architecture on the summarization task.
  - Adapted the attention mechanism in the pointer-generator architecture in Tensorflow.
  - Implemented a decoder attention mechanism to prevent repetition in the generated summary.

- *Action Recognition with Deep Learning* *Spring 2018*
  - Developed a model to recognize human actions for sequence of frames from videos using Tensorflow.
  - Built an LSTM network on top of CNN to predict an action from 11 predefined classes.

Notable Coursework

Graduate: Deep Learning, Computational Optimization, Machine Learning, Natural Language Processing, Time Series Analysis, Data Analytics, Machine Learning and Optimization

Undergraduate: Computer Programming, Digital Signal Processing, Introduction to Image Processing