# Nan Zhang

(470) 439-7958 | njz5124@psu.edu | zn1010.github.io

## EDUCATION

**Pennsylvania State University**, University Park, PA 08/2020-Present

- Ph.D. in Informatics (*Co-supervised by Prasenjit Mitra and Rui Zhang*) GPA: 3.91/4.0
- **Research Interests:** Language Model Compression, Text Summarization, Clinical NLP, Machine Learning
- **Graduate Teaching Assistant:** IST 230(Discrete Math, Fall 2020 & Spring 2021), *Instructor: Peggy Fisher*
  SRA 365(Statistics for Security and Risk Analysis, Fall 2023), *Instructor: Kevin Sylvester*

**Georgia Institute of Technology**, Atlanta, GA 08/2018-05/2020

- M.S. in Computational Science and Engineering GPA: 3.69/4.0
- **Graduate Teaching Assistant:** CS 4641/7641 (Machine Learning, Fall 2019), *Instructor: Mahdi Roozbahani*
  CS 6250 (Computer Networks, Spring 2020), *Instructor: Maria Konte*
- **Selected Coursework:** Web Search and Text Mining, Deep Learning, Bayesian Statistics, Data & Visual Analytics

**Worcester Polytechnic Institute (WPI)**, Worcester, MA 08/2013-05/2017

- BSc. in Computer Science & Industrial Engineering (double major) GPA: 3.5/4.0 (Honors with Distinction)
- Dean's List (Fall 2013, Spring 2014, Fall 2014, Fall 2016)

## PROFESSIONAL EXPERIENCE

**Research Intern**, NEC Labs America, Princeton, NJ 05/2023-08/2023

- Compressed Large Language Models (LLMs) for efficient deployment in the enterprise environment

**Software Development Intern**, Fidelity Investments, Smithfield, RI 05/2016-08/2016

- Used Spring Framework to redesign a web application and created JUnit Test Cases for its Java utility classes
- Integrated Google Analytics into a web application and conducted data analysis under Agile methodology
- Updated department calendar in the database and wrote an automated program in Java to read files and generate queries
- Utilized the FitNesse testing tool to automate acceptance testing on existing software projects

## PUBLICATION

1. **Nan Zhang**, Yusen Zhang, Wu Guo, Prasenjit Mitra, Rui Zhang. FaMeSumm: Investigating and Improving Faithfulness of Medical Summarization. Conference on Empirical Methods in Natural Language Processing (EMNLP), 2023
2. **Nan Zhang**, Shomir Wilson, Prasenjit Mitra. STAPI: An Automatic Scraper for Extracting Iterative Title-Text Structure from Web Documents. Language Resources and Evaluation Conference (LREC), 2022.

## RESEARCH EXPERIENCE

**Efficient Methods for NLP**, Advisors: Yanchi Liu & Rui Zhang & Prasenjit Mitra 05/2021-Present

- Designing pruning with fine-tuning methods to compressed Large Language Models (LLMs)

**Medical Summarization**, Advisors: Rui Zhang & Prasenjit Mitra 04/2021-12/2022

- Trained general-purpose medical summarizers that directly model faithfulness of summarization outputs to mitigate the long-lasting problem of factuality in abstractive summarization
- Designed effective methods that work on mainstream language models (e.g., BioBART, BART, T5, PEGASUS)
- Achieved state-of-the-art performance on three benchmarks of MEDIQA 2021 and a self-collected medical dialogue dataset in Chinese

**Data Analysis for Reducing Plastic Waste**, Advisor: Prasenjit Mitra 01/2022-05/2023

- Utilized state-of-the-art machine learning models to extract data from texts, tables, and figures of scholarly documents related to pyrolysis
- Designed optical character recognition (OCR) methods to better extract texts from document images in chemistry domain
- Constructed a database and a web application for users to query data and use the data extraction pipeline

**Information Extraction in Web Documents**, Advisors: Shomir Wilson & Prasenjit Mitra 08/2020-02/2021

- Trained an automatic pipeline for scraping section titles and prose text in HTML documents, which mitigates difficulties identified from existing literature
- Designed features for each text segment in HTML documents using semantic, syntactic, and visual clues
- Reached 0.97 for both weighted precision and recall on the test set, which could potentially facilitate rapid construction of large-scale section-title datasets

**Network Traffic Features Exploration to Detect Unwanted Traffic**, Advisor: Maria Konte 08/2019-12/2019

- Investigated innovative machine learning-based DDoS attack detection and mitigation solutions (by building and comparing different models) that can be deployed at the core of the Internet, within Internet exchange points (IXPs)
- Evaluated the prediction threshold to minimize false positive rate or false negative rate
- Utilized graphs and visualizations to perform data analysis of the network traffic flows to improve prediction accuracy

**BNP Paribas: Budget Automation**, *Major Qualifying Project*, Advisor: Michael Ciaraldi 10/2016-12/2016

*This is my undergraduate dissertation project, in which I worked with a business analysist to engineer an automated budgeting process for BNP's Global Markets IT Division and wrote the final paper (Grade: A).*

- Constructed the SQL database in Microsoft SQL Server which covered 44 normalized tables pulled from project tools, mapping files and front users' inputs to capture all the business terms and logic
- Designed the Python server through the Flask framework and Pandas library
- Devised the front-end via AngularJS, HTML & CSS to address various application cases like file uploading and data editing
- Addressed the identified bottlenecks within the current process and lessened the complexity of the overall process, which could potentially save hundreds of man hours

## COURSE PROJECTS

**Domain Adapation in Videos**, Instructor: Dhruv Batra 10/2019-01/2020

- Engineered Domain-Adversarial Training of Neural Networks (DANN) algorithm to apply domain adaptation for action recognition task with the souce domain as UCF-101 and target domain as HMDB-51
- Implemented a pretrained ResNet-34 to extract spatial feature of RGB frames from each video in the source domain

**Pneumonia Detection on Chest X-Rays Based on Deep Learning**, Instructor: Duen Horng Chau 01/2019-05/2019

- Trained the Convolutional Neural Network (CNN) classifier using transfer learning with Alex Net (Accuracy: 97.13%)
- Visualized and interpreted the decision-making process of CNN by using Occluding, Saliency Map, Class Activation Map, t-SNE, Activations, and Weights (Conv 1)
- Downloaded Chest X-Ray Images dataset from Kaggle containing 1,324 normal images and 3,876 pneumonia images, divided the dataset into train/val/test with ratio 80%/10%/10%, and tuned all hyperparameters via grid-search
- Constructed an interactive web application that allows users to upload their X-ray images by utlizing HTML/CSS/AngularJS on the front-end and Flask on the back-end

**Playing Atari with Deep Reinforcement Learning**, Instructor: Dhruv Batra 10/2019-11/2019

- Constructed a deep convolutional Q-network on Pong-v0 (an Atari 2600 game from OpenAI Gym) for solving Markov Decision Processes (MDPs) and utlized the RGB images of the game screen as the observations for state
- Fine tuned the model to reach a positive evaluation average score within 5 million time steps

**Variational Inference Implementation**, Instructor: Negar Kiyavash 03/2019-05/2019

- Made Bayesian inference for parameters of Gaussian mixture model and evaluated the posterior distribution via Python
- Compared the computation results with self-implemented MCMC algorithm (via R) and visualized the difference in terms of efficiency and accuracy

## TALKS AND PRESENTATIONS

- "Medical Summarization on Factuality and Readability with Large Language Models". Poster, The 10th Mid-Atlantic Student Colloquium on Speech, Language and Learning (MASC-SLL 2023). April 22, 2023
- "Transformer-based Optical Character Recognition Model for Chemistry Scholarly Papers". Poster, Penn State University ICDS Fall 2022 Symposium. October 12, 2022.

## SERVICES

- Program Committee of SUKI @ NAACL 2022
- Reviewer, ACL Rolling Review Dec. 2022
- Reviewer, EMNLP 2023
- Review Committee of CoNLL 2023

## REFERENCES

**Rui Zhang** Assistant Professor
Computer Science and Engineering Department
Pennsylvania State University
Email: rmz5227@psu.edu
Relationship: Ph.D. Advisor

**Prasenjit Mitra** Professor
College of Information Sciences and Technology
Pennsylvania State University
Email: pmitra@psu.edu
Relationship: Ph.D. Advisor