Nandan Thakur

Ph.D. Student in Computer Science @ University of Waterloo, Canada
(Semantic Scholar) (Google Scholar) (Twitter) (GitHub) (LinkedIn) (nandan.thakur@uwaterloo.ca)

Education

- **University of Waterloo** — Waterloo, Canada
  Ph.D. Student in Computer Science, Supervisor: Prof. Jimmy Lin — *Sep 2021 - Present*
- **Birla Institute of Technology and Science, Pilani (BITS Pilani)** — Goa, India
  B.E.(Hons.) Electronics & Instrumentation + Minor in Finance — *Aug 2014 - May 2018*

Employment

- **UKP Lab, Technical University of Darmstadt** — Darmstadt, Germany
  Research Assistant, Supervisors: Prof. Iryna Gurevych, Dr. Nils Reimers — *Nov 2019 - Aug 2021*
- **KNOLSKAPE** — Bengaluru, India
  Data Scientist, Manager: Mr. Chaithanya Yambari — *Sep 2018 - Oct 2019*
- **(EMBL) European Molecular Biology Laboratory** — Heidelberg, Germany
  Research Trainee, Supervisors: Dr. Toby Gibson, Dr. Manjeet Kumar — *Jun 2018 - Aug 2018*

Publications

Peer-Reviewed Conference Papers

[C3] Kexin Wang, **Nandan Thakur**, Nils Reimers, Iryna Gurevych. “GPL: Generative Pseudo Labeling for Unsupervised Domain Adaptation of Dense Retrieval” In: Proceedings of NAACL-HLT (long). 2022. [pdf] [code]. Converage: Pinecone.ai

[C2] **Nandan Thakur**, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, Iryna Gurevych. “BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models” In: Proceedings of NeurIPS Datasets and Benchmark Track (long). 2021. [pdf] [code] [leaderboard]. *Coverage*: Stanford CS224U, Open-NLP Meetup, Transformers-at-Work, ML News (Yannic Kilcher)

[C1] **Nandan Thakur**, Nils Reimers, Johannes Daxenberger, Iryna Gurevych. “Augmented SBERT: Data Augmentation Method for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks” In: Proceedings of NAACL-HLT (long). 2021. [pdf] [code] [blog]. *Coverage*: Pinecone.ai

Preprints

[P1] **Nandan Thakur**, Nils Reimers, Jimmy Lin. “Domain Adaptation for Memory-Efficient Dense Retrieval” arXiv preprint. 2022. [pdf]

Open Source

[G3] **INCOME**: https://github.com/Nthakur20/income (2022). Developer and Maintainer

[G2] **BEIR**: https://github.com/beir-cellar/beir (2021). Developer and Maintainer. Over 500+ stars

[G1] **Augmented SBERT**: https://github.com/UKPLab/sentence-transformers (2021). Developer

Research Experience

- **University of Waterloo** Waterloo, Canada
  Graduate Researcher — Supervisor: Prof. Jimmy Lin *Sep 2021 - Present*
  - Working on Domain Adaptation for Compressed Indexes. Proposed joint optimzation method improves robustness of the Binary Passage Retriever (BPR) and JPQ models across diverse zero-shot retrieval tasks and domains on the BEIR benchmark. In Progress.
  - Developing an effective sparse-retrieval baseline for zero-shot retrieval tasks on the BEIR benchmark. Currently doing a background study on implementation and evaluation of recent sparse techniques: SPLADE, TILDE, uniCOIL, DeepImpact and SPARTA.
- **UKP Lab, Technical University of Darmstadt** Darmstadt, Germany
  Research Assistant — Supervisors: Prof. Iryna Gurevych, Dr. Nils Reimers *Nov 2019 - Aug 2021*
  - Developed a heterogeneous zero-shot IR benchmark (BEIR) containing 18 datasets spanning diverse retrieval tasks and domains. Conducted quantitative analysis on out-of-distribution generalization of large pre-trained language models and traditional IR systems for robust text retrieval. [link]
  - Quantitatively analyzed cross- and bi-encoder attention networks for pairwise sentence tasks such as semantic similarity (STS). Developed a novel data augmentation technique to distill knowledge from high performing and inefficient cross-encoders to improve efficient bi-encoder performances. [link]
  - Created and open-sourced a real-world pairwise argument similarity dataset using best-worst scaling (BWS) annotation for eight controversial topics via crowdsourcing on Amazon MTurk. [link]
  - Developed a scalable entity-aware dashboard to search and cluster similar arguments for various data collections using Flask, PyTorch, Docker, VueJS and SQL. Lead an industrial project on automatic failure detection on diaper complaints with consumer argument detection and clustering. [link]
- **European Molecular Biology Laboratory, (EMBL)** Heidelberg, Germany
  Research Trainee — Advisors: Dr. Toby Gibson, Dr. Manjeet Kumar *Jun 2018 - Aug 2018*
  - Single-handedly developed a prediction toolkit using a weighted Logistic Regression (scikit-learn) model to computationally predict kinase substrate phosphorylation sites with (CAMK) protein sequences.
  - Researched heavily over dataset debiasing techniques, particularly focused on oversampling techniques such as SMOTE, nested (kxk) cross-validation to avoid overfitting during hyperparameter tuning and wilcoxon rank-sums test (α=0.05) to help us identify significant protein binding regions.
  - Conducted a thorough analysis on metrics used for evaluation of heavily-biased classification models using precision, f-measure, sensitivity and specificity along with ROC curves.

Work Experience

- **KNOLSKAPE** Bengaluru, India
  Data Scientist — Manager: Mr. Chaithanya Yambari *Sep 2018 - Oct 2019*
  - Designed and developed Krawler, an enterprise product for effectively searching a company's large messy content libraries. Developed the back-end architecture for efficient indexing. Implemented search and processing of data using Flask, Apache-Airflow, Elasticsearch and MongoDB. [link]
  - Worked on segmenting unstructured multimodal content into multiple subtopic segments. Particularly focused on unsupervised learning algorithms. Implemented and experimented with lexical (TextTiling) and semantic neural architectures for text segmentation, and conducted an error analysis.
  - Constructed an approximate content deduplication pipeline to identify near-duplicate multimodal contents. Computed hashes at scale and used Locality-sensitive hashing (LSH) and Perceptual hashing (PH) algorithms for detecting near duplicates within similar buckets for textual documents and images.

Honor and Awards

- Received University of Waterloo (UW) Graduate Scholarship for Doctoral Study in Computer Science (2021)
- BEIR: Only preprint publication to be included in teaching material in CS224U at Stanford University [link]
- Created and designed both the ELLIS NLP 2021 [link] and SustaiNLP 2021 workshop websites [link] (2021)
- Got Selected to speak for PyCon Italia titled "Extract or Replace Keywords in sentences 28x times faster than Regex - FlashText" (Cancelled due to Covid-19) (2020)
- Finalists in Technology Premier League (TPL), India's top IT strategy contest amongst fifty select corporate teams held by CIO & Leader, IT Next. (2019)
- Only UG student to receive a fully-funded Machine Learning (ML) Fellowship in EMBL, Heidelberg (2018)

Teaching Experience

- **Teaching Assistant** Waterloo, Canada
  University of Waterloo *2018 - 2019*

  1) CS 135 (Designing Functional Programs) - Fall 2021
  2) CS 136 (Elementary Algorithm Design and Data Abstraction) - Winter 2022
  3) CS 226 (Foundations of Sequential Programs) - Spring 2022

Reviewer/Program Committee

- ***CL/NLP conferences:** ACL Rolling Review: Oct-Nov (2021), Jan-Apr (2022)

Coursework

- **University of Waterloo:** CS 886: Robustness of Machine Learning (Ongoing), CS 679: Neural Networks, CS 848: Information Retrieval, CS 649: Human-Computer Interaction, CS 854: Experimental Performance Evaluation.
- **BITS Pilani:** Machine Learning, Neural Networks & Fuzzy Logic, Data Structures & Algorithms, Probability & Statistics, Linear Algebra, Econometric Methods, Discrete Mathematics.
- **Independent Study:** NLP by Deeplearning.ai (Coursera), Deep Learning for NLP (CS224d-Stanford), Machine Learning (Andrew NG), Django Introduction (Mike Hibbert), SWIRL (John Hopkins University)

Technical Skills

- **Programming:** Python, Flask, Pytorch, FastAPI, Tensorflow, VueJS, Django, JavaScript, ReactJS, R, C, C++, HTML, CSS, VBA, Advanced Excel, MATLAB, Racket, LaTeX.
- **Skills:** SQL, MongoDB, Docker, Elasticsearch, Redis, RabbitMq, Pub/Sub, Apache-Airflow, Postman.

Services

- **Machine Learning Volunteer** Bangalore, India
  Knolskape Solutions Pvt. Ltd *2018 - 2019*

  Organised a ML learning workshop for my colleagues, designed weekly assignments and took classes on Machine Learning. Implemented traditional ML models from scratch using pandas and scikit-learn.

- **Chief Coordinator, Mime Club** Goa, India
  BITS Pilani KK Birla Goa Campus *2014 - 2018*

  Led a team of 30 performers in one of the most active and popular clubs in college. Was actively involved in acting, sound mixing and creating stories for more than 12 shows over a span of 4 years for an audience of more than 2000 college students. [link]