Binh Vu

(+1) 213-269-9961 bvu687@gmail.com https://binh-vu.github.io/

Research Interests

Knowledge graph, machine learning, deep learning, and collective semi-supervised techniques to solve problems related to table understanding, information retrieval, information extraction, and question answering

Education

University of Southern California, Los Angeles, CA — Sep 2016 – May 2024
Ph.D. in Computer Science

- Thesis: Exploiting Web Tables and Knowledge Graphs for Creating Semantic Descriptions of Data Sources
- Relevant Courses: Machine Learning, Deep Learning, Representation Learning, NL Dialogue Systems
- Advisor: Professor Craig Knoblock | GPA: 3.91/4.0

HCMC University of Technology, Ho Chi Minh City, Vietnam — Sep 2010 – Jan 2015
Bachelor of Engineering in Computer Science

- Thesis: Wikipedia-based Entity Disambiguation using Deep Autoencoders
- Advisor: Professor Tru Cao | GPA: 8.73/10 (top 1%, honor program)

Projects

Critical Mineral Assessments with AI Support — Sep 2023 - Present

- Created one of the largest repositories of data about critical minerals across the world (680,000 mineral sites) by extracting and linking data from scientific publications, databases, and reports
- Developed a platform to help explore data, generate grade-tonnage models, and curate the mineral site data to assist critical mineral assessment. The system is deployed at https://minmod.isi.edu

Model Integration through Knowledge-Rich Data and Process Composition — 2018 - Aug 2020

- Developed techniques for querying and transforming TBs of heterogeneous data from separate disciplines, including geosciences, agriculture, economics, and social sciences
- Designed a language to specify transformation pipeline for automatic data conversion between different formats (GPM, GLDAS, ISRIC) and systems (Cycles, Topoflow)

Learning to Predict Cyber Attacks — Oct 2016 - Sep 2017

- Auto-crawling online sources to retrieve expertise of employees of a company, then predicting the software and hardware used in the company. A list of vulnerabilities is obtained by linking the software and hardware to the CVE database

Research Experience

University of Southern California, ISI, Marina del Rey, CA — Sep 2016 – Present
Research Assistant and then **Research Scientist**, *Center on Knowledge Graphs*

- **Semantic Table Interpretation** (for automatic data integration and knowledge graph construction)
  - Developed a novel supervised approach that uses a probabilistic graphical model to perform collective inference and outperforms state-of-the-art (SOTA) systems by 8.4% in F1 score *(accepted to WWW 2019)*
  - Developed a novel unsupervised approach, GRAMS, for linked tables. GRAMS leverages Probabilistic Soft Logic and a background knowledge graph. It outperforms SOTA methods by up to 12.6% in F_1 score *(accepted to ISWC 2021)*
  - Developed a novel distant supervised approach, GRAMS+, for general tables. GRAMS+ uses deep neural networks to help link entities and predict column types and relationships in a table. GRAMS+ outperforms SOTA approaches by 5% in F_1 score *(accepted to ISWC 2024)*
  - Developed a novel approach, GRAMS++, that is domain-independent and does not require manually labeled examples or a background knowledge graph. GRAMS++ leverages pre-trained language models and fine-tuned with distant supervision and outperforms strong LLM baselines up to 56% in F_1 score

HCMC University of Technology — Jun 2014 – Jan 2015
Undergraduate Research Assistant, *Computer Science Department*

- **Wikipedia-based Entity Disambiguation using Deep Learning**
  Using autoencoders to extract latent features of entities in Wikipedia articles for the entity linking problem

Professional Experience

Meta Inc., Menlo Park, CA — May 2021 – Aug 2021
Software Engineer Intern, *Probability Department*

- **Large-scale Feature Selection**
  - Implemented and improved mRMR, a model-free feature selection (FS) method, with new statistical tests and a novel branch-and-bound technique to improve performance and runtime in Looper, an internal auto-ML platform
  - Developed a novel grid visualization technique that simultaneously displays multiple features, enabling effective selection of important features for machine learning models
  - Reduced training data creation time by 10x by optimizing joining time, thus reducing runtime from days to hours on some huge datasets

Rakuten Inc., Tokyo, Japan — July 2015 – Apr 2016
Software Engineer, *Big Data Department*

- **Fraud Detection in ID Hijacking and Payment**
  - Developed a near real-time distributed streaming system using Apache Storm and Cassandra to analyze time-series data for fraud detection. The system is designed to run models that use related historical events up to the past 60 days to flag fraudulent transactions within seconds

Teaching Experience

University of Southern California, Los Angeles, CA — 2017, 2018
Teaching Assistant for graduate-level course INF 558 - Building Knowledge Graph

HCMC University of Technology, Ho Chi Minh City, Vietnam — 2015
Teaching Assistant for courses: Artificial Intelligence, Introduction to Programming

Technical Skills

- **Machine Learning**: PyTorch, Tensorflow, Scikit-learn, Snorkel, HuggingFace, PyTorch Lightning, PyTorch Geometric
- **Natural Language Processing**: spaCy, CoreNLP, Gensim, NLTK, HuggingFace
- **Visualization**: Matplotlib, Pyplot, ggplot2, seaborn, bokeh, plotly
- **Programming Languages**: Python, Rust, Java, Scala, C++, HTML, CSS, Javascript (Full-stack Web Developer)
- **Databases**: MySQL, Postgres, Redis, Cassandra, ElasticSearch, RocksDB
- **High Performance Computing**: Hadoop, Spark, Storm, Ray
- **Other**: Semantic Web (RDF, SPARQL, Neo4J), Docker, AWS, ReactJS

Hornors and Awards

ISI Distinguished Top-Off Fellowship — 2016

Vietnam Education Foundation Fellowship to pursue Ph.D. degree in the U.S — 2016
$54,000 for 35 selected Fellows in the whole country

Outstanding Honor Student Award — 2011 - 2024

Selected Publications

Binh Vu, Craig A. Knoblock, Basel Shbita, and Fandel Lin. 2024. *Exploiting Distant Supervision to Learn Semantic Descriptions of Tables with Overlapping Data.* In ISWC 2024 - 23th International Semantic Web Conference.

Binh Vu, Craig A. Knoblock. 2022. *SAND : A Tool for Creating Semantic Descriptions of Tabular Sources.* In European Semantic Web Conference (ESWC).

Binh Vu, Craig A. Knoblock, and Jay Pujara. 2019. *Learning Semantic Models of Data Sources Using Probabilistic Graphical Models.* In The World Wide Web Conference, pp. 1944-1953.

Binh Vu, Jay Pujara, and Craig A. Knoblock. 2019. *D-REPR: A Language for Describing and Mapping Diversely-Structured Data Sources to RDF.* In The Tenth International Conference on Knowledge Capture (K-CAP).