Binh Vu

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

Research Interests

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 **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: Craig A. Knoblock | GPA: 3.91/4.0

Ho Chi Minh City University of Technology **Sep 2010 – Jan 2015**

Bachelor of Engineering in Computer Science

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

Research Experience

Center on Knowledge Graphs, Information Sciences Institute, USC **Sep 2016 – Present**

Research Assistant

- Developed PGM-SM, a robust supervised semantic modeling approach for automated data integration. PGM-SM uses a probabilistic graphical model to collectively predict semantic descriptions of data sources and outperforms state-of-the-art (SOTA) systems by 8.4% in F_1 score (WWW 2019)
- Developed GRAMS, a novel unsupervised method leveraging Probabilistic Soft Logic to predict semantic descriptions of Wikipedia tables. GRAMS outperforms SOTA methods by up to 12.6% in F_1 score *(ISWC 2021)*
- Developed GRAMS+, a novel distant supervised approach for semantic modeling. GRAMS+ uses deep neural networks to link entities in a table and then predict column types and relationships to create the semantic description of the table. GRAMS+ outperforms SOTA approaches by 5% in F_1 score *(ISWC 2024)*
- Developed GRAMS++, the first semantic modeling approach that does not require manually labeled examples and can be applied to different domains without retraining. GRAMS++ leverages pretrained language models to learn to rank columns' types and relationships. It outperforms strong baselines up to 56% in F_1 score *(under preparation)*
- Developed D-REPR, a novel data representation language and the fastest engine to transform data to a chosen format. D-REPR can handle diversity-format datasets, including ones that require to join across multiple sources. It is used in the DARPA World Modeler project to facilitate non-materialized data exchanges between different systems and in the DARPA CriticalMAAS project to automatically ingest data into a knowledge graph. *(KCAP 2019)*
- Developed SAND, a versatile GUI for semantic modeling. SAND is designed with plugin systems for easy customization. It is used in the DARPA CriticalMAAS project to annotate and curate semantic descriptions *(ESWC 2022)*

HCMC University of Technology **Jun 2013 – Jan 2015**

Undergraduate Research Assistant

- Using autoencoders to extract latent features of entities in Wikipedia articles for the entity linking problem

Working Experience

Probabilistic Department, Meta Inc. **May 2021 – Aug 2021**

Software Engineer Intern

- Improving MRMR, a model-free feature selection (FS) method, with a novel branch-and-bound technique to scale up MRMR to handle tens of thousands of features in Looper, an internal auto-ML platform. The proposed MRMR method runs faster and better than model-based FS methods in several real use cases
- Developed Picasso, a novel feature visualization technique to display selected features in 2D images
- Improved joining time when pulling datasets to train ML models offline by 10x, thus reducing the joining time from days to hours on some huge datasets

Big Data Department, Rakuten Inc. **July 2015 – Apr 2016**

Software Engineer

- Developed a near real-time distributed streaming system for detecting fraud in ID hijacking and payment. The system is designed using Apache Storm and Cassandra to run models that use related historical events up to the past 60 days to flag fraudulent transactions within seconds
- Implemented a machine learning model for payment fraud detection

Teaching Experience

University of Southern California **2017, 2018, 2019**

Teaching Assistant, DSCI 558: Building Knowledge Graphs

- Designed and evaluated course examinations, written assignments, and weekly quizzes
- Presented several sessions of lectures to the class

Ho Chi Minh City University of Technology **2015**

Teaching Assistant, Artificial Intelligence

- Hold discussion sessions for homework and assignments
- Evaluated weekly homework and course assignments

Hornors and Awards

NSF sponsored Student Travel Awards for ISWC 2017, 2019

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 - 2014

Technical Skills

- **Machine Learning**: PyTorch, Tensorflow, Scikit-learn, Snorkel, HuggingFace, PyTorch Lightning, PyTorch Geometric
- **Natural Language Processing**: spaCy, CoreNLP, Gensim, NLTK
- **Visualization**: Matplotlib, Pyplot, ggplot2, seaborn, bokeh, plotly
- **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

Selected Publications

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, Craig A. Knoblock, Pedro Szekely, Minh Pham, and Jay Pujara. 2021. *A Graph-based Approach for Inferring Semantic Descriptions of Wikipedia Tables.* In ISWC 2021 - 20th International Semantic Web Conference.

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).

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

Professional Activities

Leader of a freelancer software development group 2012 - 2015

Organizing Volunteer in the Pacific-Asia Conference on Knowledge Discovery and Data Mining 2015

Organizing Volunteer in the Asian Conference on Machine Learning 2014

Organizing Volunteer in the Asian Conference on Information Systems 2014

References

- **Professor Craig Knoblock**
  Michael Keston Executive Director of the Information Sciences Institute
  University of Southern California, Information Science Institute, Marina del Rey, CA
  Email: knoblock@isi.edu
- **Professor Tru Cao**
  Department of Computer Science and Engineering HCMC University of Technology, Vietnam National University, Ho Chi Minh, VN
  Email: tru.cao@jvn.edu.vn