Minh Pham

Los Angeles, California
385-238-5583 minhpham1811@gmail.com Website: https://minhptx.github.io/

Research Interests

Deep learning, especially semi-supervised and active learning techniques to solve problems related to web tables and knowledge graphs: source modeling, table understanding, fact verification, anomaly detection and open-domain question answering.

Education

University of Southern California **Sep 2015 – Present**
Ph.D. Candidate in Computer Science

- Relevant Courses: Machine Learning, Building Knowledge Graphs, Advanced NLP, Representation Learning
- Advisors: Craig A. Knoblock and Muhao Chen | GPA: 3.88/4.0

Ho Chi Minh City University of Technology **Sep 2009 – Jan 2014**
Bachelor of Engineering in Computer Science

- Thesis: Wikipedia-based Entity Disambiguation | GPA: 8.5/10

Research Experience

Center on Knowledge Graphs, Information Sciences Institute, USC **Sep 2015 – Present**
Research Assistant

- Develop an approach to verify information in tabular data with table-to-text controlled text generation, natural language inference and open-domain question answering
- Developed SPADE, a robust semi-supervised error detection approach. SPADE leverages Probabilistic Soft Logic and deep neural networks to detect errors in tables and yield a 10% improvement in accuracy when compared to the state-of-the-art systems *(accepted to IJCAI 2021)*
- Developed UDATA, a novel unsupervised data transformation system. UDATA requires no labeled examples and can transform string values between different formats with 80% accuracy *(accepted to IEEE BigData 2019)*
- Developed DSL, a learning-based method for domain-independent semantic labeling. A pre-trained DSL model can achieve an average of 83% in MRR ranking score across different domains without fine-tuning *(accepted to ISWC 2016)*

Artificial Intelligence and Language Lab, Nuance Communications Inc. **Jun 2018 - Aug 2018**
Research Intern

- Proposed an unsupervised method for an internal entity resolution project using Probabilistic Soft Logic

John von Neumann Institute, Vietnam National University **Sep 2014 – May 2015**
Research Assistant

- Developed an entity linking method that outperforms state-of-the-art systems by 3% in F1-score by combining candidate searching and rule-based coreference resolution

Projects

Model Integration through Knowledge-Rich Data and Process Composition **Sep 2017 – Dec 2019**

- Designed a data-streaming pipeline to process TBs of heterogeneous data from separate disciplines, including geosciences, agriculture, economics, and social sciences
- Constructed 10 transformation adapters and 5 transformation pipelines to convert data between different formats (GPM, GLDAS, ISRIC) and systems (Cycles, Topoflow)

Probabilistic Representation of Intent Commitments to Ensure Software Survival **Jan 2016 – Sep 2017**

- Developed a learning-based system to distinguish different signals in sensor data (based on DSL)

Professional Experience

DataFirst JSC Co. **Feb 2015 - Aug 2015**
Research Programmer

- Extracted real estates' information from millions of Vietnamese online listings with high accuracy for market analysis

East Agile **Jun 2014 – Aug 2014**
Software Engineer Intern

- Developed and maintained an in-house video sharing platform using Ruby on Rails, CoffeeScript, HTML5 and CSS.

Teaching Experience

University of Southern California **2016 - 2021**
Teaching Assistant, DSCI 558: Building Knowledge Graphs

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

Ho Chi Minh City University of Technology **2013 - 2015**
Teaching Assistant, Artificial Intelligence

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

Selected Publications

Pham, M., Knoblock, C., Chen, M., Vu, B. and Pujara, J., *SPADE: A Semi-supervised Probabilistic Approach for Detecting Errors in Tables.* In 30th International Joint Conference on Artificial Intelligence (IJCAI 2021).

Pham, M., Knoblock, C. and Pujara, J., *Learning Data Transformation with Minimal User Effort.* In 2019 IEEE International Conference on Big Data (IEEE BigData 2019).

Pham, M., Alse, S., Knoblock, C., and Szekely, P., *Semantic Labeling: A Domain-Independent Approach.* In 15th International Semantic Web Conference (ISWC 2016).

Awards and Scholarships

Best Paper Award, ISI Graduate Student Symposium, University of Southern California **2019**
Title: Learning Data Transformations with Minimal User Effort

Vietnam Education Foundation (VEF) Fellowship for Ph.D. study in US **2015**
$54,000 for 35 selected Fellows in the whole country

Leadership / Extracurricular

Vietnam Education Foundation Fellows and Scholars Association (VEFFA) **2016 - 2017**
Board of Executives

- Organized VEFFA annual conferences and events.
- Organized mock interviews for more than 60 Vietnam Education Foundation (VEF) Scholarship applicants.
- Led a group of 20 mentors to support VEF Scholarship applicants in preparing their applications.

PAKDD 2015 Conference & ACML 2014 Conference **2014 - 2015**
Website Administrator and Volunteer

- Designed and managed conference website.
- Monitored presentation sessions in the conference.

Online Courses

Machine Learning Data Lifecycle in Production (Coursera) **2021**
Introduction to Machine Learning in Production (Coursera) **2021**

Technical Skills

- **Machine Learning**: PyTorch, Tensorflow, Keras, Scikit-learn, Snorkel, HuggingFace
- **Natural Language Processing**: spaCy, CoreNLP, Gensim, NLTK, FairSeq, OpenNMT
- **Visualization**: Matplotlib, Pyplot, ggplot2, seaborn, bokeh, plotly
- **Languages**: Python, Java, Scala, C++, SQL
- **Semantic Web**: RDF, Turtle, SPARQL
- **High Performance Computing**: Dask, Spark
- **Databases**: ElasticSearch, MongoDB
- **Web Development**: Django, Ruby On Rails, Flask
- **Others**: Docker, AWS