Lam Thanh Do

louisdo15111999@gmail.com · https://github.com/louisdo · (+84)902131780

EDUCATION

HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY *Fall 2018 - Spring 2023*
BSc. in Computer Science

PUBLICATIONS

1. **Lam Thanh Do**, Pritom Saha Akash, Kevin Chen-Chuan Chang. Unsupervised Open-domain Keyphrase Generation. The 61st Annual Meeting of the Association for Computational Linguistics (ACL). 2023.

RESEARCH EXPERIENCE

Forward Data Lab - University of Illinois at Urbana-Champaign *Jan 2021 - Present*
Research Intern

- Supervisor: Prof. Kevin Chen-Chuan Chang
- Research topics: Unsupervised Keyphrase Generation, LLMs Scheduling.

INDUSTRIAL EXPERIENCE

CAZOODLE INC. *June 2023 - Present*
Software Engineer

- In charge of the Academic Online project, a project that collects and combines scholarly data from various sources on the internet, to aid users in understanding the landscape of research.
- Lead a team of 8 members. In charge of designing the data collection and processing procedure.
- Work on both the front-end (ReactJS) and back-end (Python).

ICOMM VIETNAM *December 2022 - June 2023*
NLP Engineer

- Take part in the project of **ICGPT**, a **generative AI** that synthesizes reports in the politics domain, alongside 2 members. The aim of the project is to collect information from various sources and then synthesize this information into a single report. Developed the English version of ICGPT using the pre-trained model **Flan-T5**.
- Take part in the project **Icomm Chatbot**, a **generative conversational AI** in domains of healthcare services and banking. The aim of the project is a conversational AI that can recommend services to users. Developed the **retrieval model** that obtains information based on the conversation history. Took part in building the generative model using the pretrained model **BARTPho**.

VIETTEL GROUP (VIETTEL Data Governance Department) *January 2022 - October 2022*
Data Scientist (Insource, Part-time)

- Take part in developing the **Viettel Machine Learning Platform**, a product which aids organization in solving data analysis as well as machine learning tasks**.**
- In charge of the project **Viettel Recommendation Engine** (vRE). This project aims at assisting businesses in the deployment and management of recommendation systems. In charge of designing the software architecture and the algorithms that vRE uses. Implemented the algorithm using **spark** and integrate with the web app.

PIXTA Inc. (PIXTA Vietnam) *June 2020 - April 2021*
Junior Research Engineer (Part-time)

- Developed version 2 of the **Learning-to-rank** project alongside 2 members, the project aims to enhance the quality of Pixta's **search engine**. Improved the data through the **feature selection** and **feature engineering** phase, resulting in improving the model's NDCG score by 3%. Designed a **diversification algorithm** that runs in linear time and deployed it to Elasticsearch. Developed an **online-learning system** that automatically acquires data, trains the

model and uploads that model to staging. In the first A/B test, the project managed to increase Click-through rate by 15% and increase Conversion-rate by 4%.

- Developed version 2.1 of the **Tag Suggestion Project**. Designed **an algorithm that makes use of user feedback** to increase precision of the model by 3.5% and increase F1 by 1.5% in comparison to version 2.0.

PIXTA Inc. (PIXTA Vietnam) *September 2019 - June 2020*

Research Engineer Intern

- Developed version 3 of the **Pixta Highscore Project** alongside 4 members. **Pixta Highscore Project** is a project with the purpose of automatic ranking tags among user-annotated tags; designed and implemented a model based on **Transformer** that achieved an F1-score of 52%. The project got deployed to production.
- Built a package named **pixtaresearch** consisting of frequently used functions and utility classes to improve team members' code. This package is used by all members in the research team and helps save a remarkable amount of time.
- Developed version 2.0 of the **Tag Suggestion Project** alongside 1 member, a project aiming at assisting users in annotating tags to images in a fast, convenient way and professionally; developed a model using Resnet50 and a **customized binary cross entropy loss function** with labeling uncertainty, the model achieves an F1-score of 54%, outperforms by more than twice in comparison to the 20% F1-score of the previous version on the same dataset and got deployed to production.

RubikAI *May 2019 - August 2019*

Computer Vision Trainee (RubikTalent Program)

- Learned Machine Learning, Deep Learning, Image Processing concepts and techniques.
- Learned to use frameworks like Tensorflow to build full-fledged programs.

TECHNICAL SKILL

Programming Languages: Proficient in Python; comfortable with SQL, Javascripts; familiar with C/C++, Java.
Frameworks: Proficient in using pytorch, numpy, pandas, pyspark. Comfortable with ReactJS.
Mathematics Skill: Strong knowledge of Linear Algebra, Calculus, Statistics, Data Structure and Algorithm.
Foreign Language: English with professional working proficiency.

SIDE PROJECTS

These-Images-Does-Not-Exist - A program that generates images.

- Collect images from Flickr and train a **Deep Convolutional Generative Adversarial Network** to generate images of different categories.

Narratives Generator - A program that takes in a list of suggested words and outputs a full narrative.

- Uses the language model **Transformer** as well as other text processing methods such as **regex**.

MoviesSearch - A **search engine** for the IMDb movie dataset built in **Python**

- The program makes use of the **TF-IDF** ranking methods and sparse matrix representation.

IDEX - An open-source Vietnamese ID card **OCR program** written in Python

- The program scans ID cards from an image and extracts desired information using Image Processing and Deep Learning techniques such as **Canny Edge detection**, **Contouring**, **RetinaNet**.