Zhiyu (Edward) Liang

420 Temple St, New Haven, CT 06511
475-434-8259 zhiyu.liang@yale.edu LinkedIn Page

Education

Yale University **Aug. 2021 – May 2022**
Master of Science in Computer Science *New Haven, CT*

University of Toronto **Sep. 2015 – Dec. 2019**
Honours Bachelor of Science in Computer Science; ***GPA: 3.84/4.00*** *Toronto, ON*
Artificial Intelligence and Computer Vision Track

Technical Skills

Languages: Python, Swift, C/C++, Java, Shell, SQL, MATLAB, JavaScript, HTML, CSS, Markdown
Technologies: Docker, Google Cloud, Node.js, Bootstrap, iOS, Linux, Git, RegEx
Machine Learning: PyTorch, TensorFlow, Keras, NumPy, OpenCV, Fairseq, Tensor2Tensor, Scikit-Learn, Pandas

Experience

Qualcomm **Feb. 2020 – Jul. 2021, May 2018 – Apr. 2019**
Machine Learning Software Engineer *Toronto, ON*

- Developed 2 algorithms to reduce the size of **Transformer models** by **4x** while preserving **97.7%** model accuracy in **machine translations**.
- Implemented a Python script to translate internal model representations into ONNX format using Regular Expression, which accelerated the processing time **from 2 hours to 5 seconds**.
- Proposed a deep learning based programming language translation algorithm, implemented the pipeline in PyTorch from data collection of 550,000 training pairs, preprocessing, training and evaluation.
- Jointly developed the Quantization-Friendly MobileNet, won the **1st prize in 2018 IEEE Low Power Image Recognition Challenge** and published a **NeurIPS Workshop paper**.

Vector Institute **May 2019 – Dec. 2019**
NLP Research Intern *Toronto, ON*

- Crawled and preprocessed large-scale high-quality text data of **12 billion word tokens**.
- Implemented the **GPT-2 model** in Tensor2Tensor and training scripts for **distributed training** setup, available for hundreds of researchers and sponsors to use.
- Profiled distributed training performance of GPT-2 on **256 GPUs**, which helped decide the multimillion-dollar purchase of additional training hardware.

Projects

FlowerWiki App | *Swift, Xcode, CoreML, Wikipedia API, MVC Design Pattern* **Aug. 2021**

- Developed an iOS app using Swift which **recognizes 102 flower types on-device** from the photo taken by the user, and presents a brief description and a picture of the flower from Wikipedia.
- Converted a pre-trained Caffe model for flower recognition into mlmodel format required by CoreML.
- Implemented the pipeline from photo capturing to preparing and feeding the photo to the model for prediction.
- Fetched flower descriptions and sample photos from Wikipedia API using HTTP requests.

To-Do List App | *Swift, Xcode, Realm, MVC Design Pattern* **May 2021**

- Developed an iOS app using Swift and Realm to help users keep track of their to-dos.
- Designed GUI for the main category page and to-do items page in Xcode storyboard.
- Implemented creating, reading, updating and deleting (CRUD) of user data with **Realm by MongoDB**.
- Integrated swipe-to-delete function using SwipeCellKit.

FaceBlock App | *Tensorflow, Java, Android Studio* | *2018 Qualcomm Hack Mobile Winner* **Jun. 2018**

- Developed an Android app using Java and TensorFlow to protect people's privacy in live video streams by detecting, tracking and replacing unwanted faces with a selected emoji in real time on Samsung Galaxy S9.
- **Won 1st place in 2018 Qualcomm Hack Mobile Hackathon** out of 60+ teams and 250+ participants.

Publication

Low Power Inference for On-Device Visual Recognition with a Quantization-Friendly Solution
Chen Feng, Tao Sheng, **Zhiyu Liang**, Shaojie Zhuo, Xiaopeng Zhang, Liang Shen, et al
Neural Information Processing Systems 2018 MLPCD 2