Zetian (Neal) Wu

+1-(443)-630-2430 | zwu49@jhu.edu | neal-ztwu.github.io

Education

Johns Hopkins University — Maryland, United States
MSE in Data Science — *Jan. 2020 – Dec. 2021*

Zhejiang University — Zhejiang, China
BS in Physics | Minor in Finance — *Sept. 2015 – Jun. 2019*

Research Experience

Research Assistant — Jun. 2020 – Present
with Prof. João Sedoc, Prof. Lyle Ungar — *NYU/UPenn, United States*

Investigation in Explainabiliby Method for Lexicon creation (*Plan to NAACL 2022*)

- **Problem**: Trying to understand how deep network models take advantage of language data and how we are supposed to interpret models so that they can aid us in explaining the world.
- **Techniques**: Used diverse methods to compute scores for all the words in each sentences, i.e. creation at instance level, and evaluated the results from the human interpretation side. Trying to find out the reason why unexpected results came to birth, and to implement better methods based on current algorithms and observations.

Lexicon Creation for Interpretable NLP Models

- **Problem**: Compared the quality of lexica generated from different model-method pairs and made conclusions concerning explanability features of different models and methods accordingly.
- **Techniques**: Built FFN, SVM, RoBERTa, DistilBERT models and implemented lexicon generation methods including single token importance, masking and Partition Shap to create lexicon. Evaluated lexica in terms of generalization ability and human annotation.
- **Results**: Context-sensitive models generalized better to different datasets in similar domain while lexica created from them on the contrary performed worse compared to that created from context-oblivious models.

Research Assistant — Mar. 2021 – Present
with Prof. Louis-Philippe Morency — *CMU, United States*

Multimodal Multitask Training (*In progress*)

- **Problem**: Trying to build a multitask model which can handle diverse modalities and tasks concurrently and achieve comparable performance with state-of-the-art single task models.

MultiBench: Multiscale Benchmarks for Multimodal Representation Learning

- **Problem**: Built a benchmark for multimodal fusion models and datasets.
- **Techniques**: Implemented several multimodal fusion methods including early/late fusion, LRTF, Mutual Information Matrix, CCA, RefNet, MFM and RMFE. Built a universal codebase to train and evaluate each model under different datasets on according metrics and robustness.

Research Assistant — Apr. 2020 – Jan. 2021
Center for Language and Speech Processing, with Prof. Benjamin Van Durme — *JHU, United States*

Span Identification and Representation for Information Extraction

- **Problem**: Formulated entity mention detection problem under partially annotated datasets.
- **Techniques**: Built an LSTM-based model to detect spans by conditioning on given spans. Introduced a ranking loss to rank gold spans higher while not fully ablating unlabelled spans. Took the SpanBERT-based coreference model as span proposal model to detect entity mentions.
- **Results**: Achieved recall above 0.9 and F1 score above 0.8 when conducting few-shot finetuning.

Research Assistant — Apr. 2018 – Aug. 2019
Intelligent Computing & System Lab, with Prof. Qinming He — *ZJU, China*

Anti-fraud Model for New Financial Leasing Services
(*Top Prize in China Collegiate Computing Contest-AI Innovation Contest*)

- **Problem**: Built an anti-fraud model for online financial leasing services.
- **Techniques**: Constructed two kinds of features: statistical features from Bipartite Graph and the node representations from Unipartite Graph using DeepWalk. Implemented DeepFM as the supervised learning model.
- **Results**: Increased AUC by 6% compared to the best baseline (GBDT Tree with Logistic Regression).

Interactive Rare-Category-of-Interest Mining from Large Datasets

- **Problem**: Built a web crawler for data collecting and a CNN-based feature extractor to construct a real audio dataset (Birdcall) along with a numerical dataset (Medicine) for performance evaluation.
- **Techniques**: Implemented a Rare Category Detection (RCD) model using a combined method of offline phase inference and high-level knowledge abstractions, reducing the time complexity of query answering from quadratic to logarithmic. Built a Rare Category Exploration (RCE) model using a collaborative-reconstruction approach.
- **Results**: Obtained at least 11.75% improvement in accuracy compared with baseline algorithms including kNN, Interleave, NNDM, Clover, and FRANK.

Work Experience

Research Intern Apr. 2021 – present

Microsoft Research Asia *China*

Style-Specific Melody Generation in an Unsupervised Way (*In progress*)

- Trying to use dimension reduction methods and appropriate evaluation methods to select midi features and label data points using clustering methods.
- Trying to built conditioned generation models such as vq-vae to generate style-fixed midis.

Machine Learning Engineer Aug. 2019 – Mar. 2020

Hangzhou Enjoymusic Technology Co. Ltd. *China*

- Built a sequence-to-sequence model for music style transferring using TransformerXL and Discriminator.
- Formulated automatic music piece generation problem as a conditional sequence generation task that decodes MIDI sequence from drum beats, and modelled with VAE architecture.
- Refactored Typescript Midi-me codes using Python for integration with our own platform and application.

Publications

Zetian Wu*, Yilin Geng*, Roshan Santhosh, Tejas Srivastava, Lyle Ungar and João Sedoc. Lexicon Creation for Interpretable NLP models. Submitted to *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Yufan Chen, Peter Wu, Michelle A Lee, Yuke Zhu, Russ Salakhutdinov, and Louis-Philippe Morency. Multibench: Multiscale benchmarks for multimodal representation learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021

Zhenguang Liu, Sihao Hu, Yifang Yin, Jianhai Chen, Kevin Chiew, Luming Zhang, and Zetian Wu. Interactive rare-category-of-interest mining from large datasets. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):4965–4972, Apr. 2020

Skills and Additional Information

Programming/Framework: Python, PyTorch, TensorFlow, AllenNLP, Linux, C/C++, MATLAB, R, SQL
Awards: Top Prize in China Collegiate Computing Contest-AI Innovation Contest, Honorable Award in COMAP
Honors: 2nd Level in Training Plan of the National Basic Subject Top-notch Talent Scholarship