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*

Human-interpretable Lexicon creation from NLP Models (*In progress*)

- **Implementation**: Used diverse lexicon generation methods to give scores for all the words in each sentences, i.e. creation at instance level.
- **Analysis**: Trying to evaluate the results from the human interpretation side, and to find out the reason why some unusual situation happens, and to give mathematical explanations.
- **Method Improvement**: Come up with a effective method based on former findings.

Lexicon Creation for Interpretable NLP Models (*Paper submitted to ACL 2022*)

- **Data Cleaning**: Removed non-English words and HTML strings in Yelp Reviews, Amazon Reviews and NRC dataset, and conducted down sampling to balance labels.
- **Implementation**: Built FFN, SVM, RoBERTa, DistilBERT models and implemented lexicon generation methods including single token importance, masking and Partition Shap to create lexicon.
- **Evaluation**: Evaluated lexica in terms of generalization ability and human annotation, and conducted paired t-test to examine the significance of performance difference between models and lexica and among diverse lexica.

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

Multimodal-Collaborative Pretraining (*In progress*)

- **Feature Extraction**: Applied R-CNN and Audio Spectrogram Transformer separately to extract image and audio features, and conducted hierarchical K-means cluster to tokenize audio sequences.
- **Implementation**: Trying to build BERT-related models using co-attention and transformer-based models to deal with multiple modalities and to compare their results as well as with former MultiBench results.

MultiBench: Multiscale Benchmarks for Multimodal Representation Learning
(*Paper accepted by NeurIPS 2021 Track Datasets and Benchmarks*)

- **Data Processing**: Implemented data reader for MOSI, MOSEI and MM-IMDB dataset.
- **Model Training**: Implemented several multimodal fusion methods including early/late fusion, LRTF, Mutual Information Matrix, CCA, RefNet, MFM and RMFE.
- **Pipeline Design**: Built a universal codebase to train and evaluate each model on different datasets including robustness evaluation.

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

- **Task Design**: Formulated entity mention detection problem under partially annotated datasets as a span ranking task, where a ranking loss is enforced to rank gold spans higher while not fully ablating unlabelled spans.
- **Implementation**: Built an LSTM-based model to detect spans by conditioning on given spans, supporting extraction tasks such as event extraction.
- **Model Improvement**: Investigated taking the SpanBERT-based coreference model as span proposal model to detect entity mentions, achieving recall above 0.9 and F1 score above 0.8 when finetuned with only a few training examples.

Research Assistant Apr. 2018 – Aug. 2019

Intelligent Computing & System Lab, with Prof. Qinming He *Zhejiang University, China*

Anti-fraud Model for New Financial Leasing Services

(*Top Prize in China Collegiate Computing Contest-AI Innovation Contest*)

- **Feature Extraction**: Constructed two kinds of features: one is obtained from Bipartite Graph as statistical features and the other is extracted from Unipartite Graph using DeepWalk model as node embeddings.
- **Implementation**: Built supervised learning model (DeepFM), increasing the anti-fraud ability of the new financial leasing services by 6% on AUC.

Interactive Rare-Category-of-Interest Mining from Large Datasets (*Paper accepted by AAAI 2020*)

- **Feature Extraction**: 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.
- **Implementation**: 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; and built a Rare Category Exploration (RCE) model using a collaborative-reconstruction based approach, and compared our model with baseline algorithms including kNN, Interleave, NNDM, Clover, and FRANK, resulting in at least 11.75% improvement in accuracy.

Work Experience

Research Intern Apr. 2021 – present

Microsoft Research Asia *China*

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

- **Feature Engineering**: Trying to use dimension reduction methods and appropriate evaluation methods to select midi features and label data points using clustering methods.
- **Implementation**: 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.

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