Yufei **Li**

Computer Science · Machine Learning Researcher

(+1) 858-291-2678 | yxl190090@utdallas.edu | yul091.github.io/liyufei.me/ | yul091 | @CavMonkey

Research Interests

Research at the intersection of natural language processing (NLP) and data mining. Interested in the application of NLP in software engineering (SE) and recommender systems, as well as model uncertainty measurements.

Education

The University of Texas at Dallas — *Dallas, TX*

Ph.D. student in Computer Science — *2020 - Present*

Artificial Intelligence Group, CSE; Advisor: Dr. Wei Yang

University of California, San Diego — *San Diego, CA*

M.S. in Electrical and Computer Engineering — *2018 - 2020*

Thesis: Application of A* and Genetic Algorithm in TSP Path Planning Problem; Advisor: Dr. Farinaz Koushanfar

Xi'an Jiaotong University — *Xi'an, China*

B.S. in Mechanical Engineering — *2014 - 2018*

Thesis: Aircraft fuel tank oil sloshing simulation based on SPH method and center of gravity distribution analysis; Advisor: Dr. Shuai Zheng

Publications

SHARE: a System for Hierarchical Assistive Recipe Editing.

Li, S., **Li, Y.**, Ni, J., McAuley, J. | *arXiv Computing Research Repository preprint:2105.08185 (CoRR 2021)* [PDF]

- Identified a controllable recipe editing task: adapting a base recipe to satisfy a user-specified dietary constraint to assist home cooks.
- Proposed a hierarchical framework that performs simultaneous ingredient substitution via a Set Transformer, followed by conditional instruction generation using a Copy Attention network.

Estimating Predictive Uncertainty Under Program Data Distribution Shift.

Li, Y., Chen, S., Yang, W. | *arXiv Computing Research Repository preprint:2107.10989 (CoRR 2021)* [PDF]

- Defined three real-world program distribution shifts based on practical software development scenarios.
- Investigated the effectiveness of existing uncertainty measurements under program distribution shifts and provided a large-scale benchmark for their performance.
- Analyzed each method's pros and cons from a logic design perspective to inspire a domain-specific uncertainty design.

GLIB: Towards Automated Test Oracle for Graphically-Rich Applications.

Chen, K.*, **Li, Y.***, Chen, Y., Fan, C., Hu, Z., Yang, W. | *29th Foundations of Software Engineering (FSE 2021)* [PDF]

- Proposed a code-based data augmentation approach for generating UI glitch images on game apps.
- Detected UI glitches via a CNN model and localized the glitch area using a saliency map to facilitate bug fixing.

Work Experience

NEC Laboratories America, Inc. — *Princeton, NJ*

Research Intern, PyTorch — *Jun 2021 - Aug 2021*

- Annotated name entities and relations with regular expression rules on CVE texts for distant supervision.
- Incorporated the pre-trained GPT-2 backbone into a sequence labeling framework for joint entity & relation extraction.
- Proposed a bootstrap training procedure for denoising distant labels and selecting high-quality instances.

SeekTruth Scientific and Technical Corporation — *Beijing, China*

Research Intern, TensorFlow — *Jul 2019 - Sep 2019*

- Boosted an adaptive discrimination definition model for objection detection.
- Designed a lightweight CNN model for identifying the direction of videos.
- Built a joint key point & pose recognition model for character detection.

Projects

Enhance the Reliability of Deep Networks on Out-of-distribution Data

Dallas, TX

NLP & CV, PyTorch

Sep 2021 - Present

- Perturbing the text inputs from unknown distribution via synonym substitutions to yield trustworthy predictions.
- Boosting the CNN model confidence of test inputs through label-preserving image transformations.

Low-Resource Joint Entity and Relation Tagging with Distant Supervision

Dallas, TX

NLP & Data Mining, PyTorch

Sep 2021 - Present

- Identifying the instance quality as the explainability of models' position attention distribution.
- Proposing a data redistribution schema to dynamically train on weakly labeled data.

Assessing the Reusability of Pre-trained Code Embeddings

Dallas, TX

NLP & SE, PyTorch

Sep 2020 - May 2021

- Proposed a low-cost offline metric for evaluating the generalizability of code embedding in SE downstream tasks.
- Patched the generalizability of existing pre-trained embedding based on the semantic metamorphic relationship.

Rethink Negative Sampling in Bayesian Personalized Ranking

San Diego, CA

Data Mining & Recommender Systems, PyTorch

Nov 2019 - Jun 2020

- Analyzed one limitation of the popularity-based sampling scheme in terms of non-uniform negative sampling bias.
- Corrected the bias and designed related negative sampling distributions to boost the Bayesian personalized ranking (BPR).

Automatic Delivery Vehicle Design

San Diego, CA

Path Planning, Python & MATLAB

Mar 2019 - Jun 2019

- Incorporated the Courier and TSP travel agent problems into designing autonomous delivery vehicles.
- Designed a path planning algorithm by incorporating the A* heuristic rules into the genetic evolution.

How Cute is My Pet? Pet Adoption Speed Prediction

San Diego, CA

NLP & CV, TensorFlow

Jan 2019 - Mar 2019

- Parsed and tokenized the description text of pets, encoded the semantic features with BiLSTM.
- Designed a joint framework to fuse the image feature, semantic feature, and other properties for pet popularity prediction.

Awards

VEX Robotics International Competitions

Path Planning, C & C++

- Excellent award and runner-up in VEX Robotics World Championship (RECF) 2017, Louisville, KY, US.
- Excellent award and runner-up in VEX Robotics Asia Open 2016, Beijing, China.
- First-class honor in VEX Robotics China Open 2016, Xi'an, China.

Scholarship Awards

Personal

- National encouragement scholarship 2015-2017, top10 of 300+ applicants.