Yangcheng Gu

Tel: +86-135-8553-0910
E-mail: guyc19@mails.tsinghua.edu.cn
Webpage: https://ferv3455.github.io/

Education

Tsinghua University Sept. 2019-Jun. 2024 (expected)
Bachelor of Engineering in Software Engineering
Overall GPA: 3.88/4.00
Standardization Examination: TOEFL: 110 (R30+L29+S23+W28); GRE: 325 (V155+Q170) +AW 4.0

Paper

Li, X.; Yang, P.; **Gu, Y.**; Zhan, X.; Wang, T.; Xu, M.; and Xu, C. "Deep Active Learning with Noise Stability." Manuscript submitted to the AAAI-24 Conference for publication.

Research Experience

Computational Biology Department, Carnegie Mellon University
Advisor: Min Xu, Associated Professor Jul. 2023-Present

Deep Active Learning with Noise Stability

- Participated in proposing and developing a novel algorithm leveraging noise stability to estimate prediction uncertainty for unlabeled data, in which random noises are added to model parameters, and the consequent perturbation of output is regarded as model prediction stability
- Conducted extensive theoretical analysis using the small Gaussian noise theory, and proved that the novel algorithm favors subsets with large and diversified gradients
- Reproduced the algorithm of Active Learning by Feature Mixing (ALFA-Mix) based on our architecture, implemented a series of validation experiments on various standard machine learning datasets, and evaluated the proposed algorithm against it, producing results in favor of our method generally
- Proposed an improvement method for the clustering algorithm used in our method, yielding a training accuracy increased by 1-2% in certain training scenarios
- Assessed the performance of baseline algorithms and the new algorithm on datasets with the Average Rank method, proving that our proposed algorithm ranked 1st generally and not worse than 2nd in all tested cases

Bi-level Optimization for Inductive Transfer Learning

- Proposed a model pretraining method in which distinct sample weights are assigned to source data with a neural network learned with the DARTS method so that the pretrained model can achieve better performance when transferred to a specific, differently distributed dataset
- Located and fixed errors in the original code to ensure the successful reproduction of results for typical transfer learning tasks, and refactored the code to enable unified usage for different tasks
- Achieved satisfactory sample weight results by using subsets of MNIST, CIFAR-10 and CIFAR-100 datasets, which accurately reflected the correlation between source classes and target data
- Assessed the performance of the proposed method in specific scenarios where hybrid datasets with samples from multiple domains are used as source datasets, and produced an evident improvement of about 3% in accuracy with a limited target training dataset compared to baseline methods
- Combined the proposed method with SimCLR self-supervised learning architecture, and looked into the prospect of adopting sample weighting in the domain of self-supervised deep learning by conducting experiments with the CIFAR-10 dataset
- Explored the application of the proposed method to specialized datasets, including various medical and clinical datasets

School of Software, Tsinghua University
Log Data Encoding for Efficient Storage in Apache IoTDB
Advisor: Shaoxu Song, Associated Professor Oct. 2022-May. 2023

- Raised the concept of advanced character operations, and constructed the edit path of such operations from one string to another based on dynamic programming, which could be used as encoding information
- Added weights to different characters in the algorithm based on values correlated to frequencies such as Huffman codes, considering that commonly used characters could be compressed more efficiently
- Maintained the source string pool with a sliding window for selecting the one with the shortest distance
- Classified the assortment of log data according to data formats by applying clustering algorithms, minimizing necessary data storage
- Handled the temporal optimization of the algorithm by adopting an approximate string distance algorithm, Cosine Distance based on Q-grams, with linear time complexity

Yangcheng Gu

Tel: +86-135-8553-0910
E-mail: guyc19@mails.tsinghua.edu.cn
Webpage: https://ferv3455.github.io/

Developing a Network Security Data Management System
Advisor: Hai Wang, Associate Research Fellow Oct. 2023-Present

- Designed and improved the user login interface and its functions, including registration, login, logout, personal information management, and resetting the password
- Participated in the collection and maintenance of the attack behavior dataset, and the generation and maintenance of the provenance graph dataset
- Implemented the algorithm details page, including the algorithm introduction, model assessment and code submission, and designed the navigation bar of the web page
- Designed test cases in the functional and end-to-end tests of the web application

Visual Perception and Understanding of Three-dimensional Scenes: Action Recognition
Advisor: Yue Gao, Associate Professor Oct. 2021-Sept. 2022

- Aimed to build a DVS action dataset and implement methods for action recognition through DVS video
- Collected the largest DVS dataset of 50 actions with 2330 segments
- Tuned the existing model for the DVS dataset to achieve an improvement in the classification accuracy
- Constructed a new model based on the GNN framework and fitted it to a simplified subset of the dataset
- Attempted to extend the dataset, and tried to optimize the network model based on the GNN framework in order to improve the classification accuracy

Internship Experience

Risk Control R&D Department, Beijing Jingdong Century Trading Co., Ltd.
Supervisor: Nan Lu, Algorithm Application Engineer Jun. 2023-Aug. 2023

Click Farming Detection based on Comment Content and Language Model

- Learned key technical means of click farming detection, and acquired necessary knowledge concerning the field of big data including Hadoop, Hive, Spark and Oozie by participating in practical projects
- Mastered the principle and application of click farming detection models based on comment content and BERT language model and on customer and seller behavior on the big data platform provided by JD
- Fitted the existing pretrained click farming detection model to a comment dataset of another business scope, mainly by modifying the data processing procedure and tuning configuration parameters

Click Farming Detection based on the Similarity of Comment Pictures

- Mastered the implementation and applications of various existing image similarity algorithms, evaluated their performance on real comment image datasets, and studied the feasibility of identifying suspected click farming according to image similarity
- Put forward a special image similarity algorithm based on measuring the abnormal local identity of images with statistic methods and the feature matching algorithm, which yielded an average accuracy of over 95% on experimental synthetic datasets and over 98% on real datasets
- Independently developed a click farming detection method with the proposed image similarity algorithm and the comment image dataset, which could detect almost 100% of the tested anomalous image groups
- Completed the optimization and environment configuration of the algorithm and deployed it as a scheduled task in the production environment successfully, processing up to 300,000 images per day

Project Experience

Design and Implement a MarsOJ Youth Algorithm Learning Platform

- Implemented the real-time communication framework in the frontend project, designed the login registration page, the home page, the battle logic of the competition system, and the personal information page, and realized the overall page optimization
- Participated in the collaborative development and version control of the system by maintaining and contributing to Git repositories of the frontend and backend projects on the CODING platform

Design a Campus Information Release Platform

- Designed the database and the entire backend framework of the system with the Python Django framework
- Implemented most functions of the mobile application using Android UI, including browsing post lists, creating new posts and the notification system

Others

- **IT:** Python, C/C++, Java, JavaScript, HTML/CSS, PyTorch, Scikit-Learn, SQL, Linux
- **Personal Hobbies:** Reading, karting, billiards, soccer, swimming, board games, etc.