Zihao Wang

zw1074@nyu.edu | (551)225-9955 | 204 10th St, Apt. 314, NJ 07302 | https://zw1074.github.io

Educations

New York University, Courant Institute of Mathematics **New York, NY**
*Master of Science in **Data Science (3.80/4.0)*** *May 2017*
Nanjing University **Nanjing, China**
*Bachelor of Science in **Computational Mathematics (3.50/4.0)*** *Jun 2015*

Related Courses
Machine Learning, Probability Theory, Stochastic Processes, Statistical Inference, Business Understanding for Data Science, Big Data, Optimization Theory, Deep Learning, NLP, Advanced Python, Fundamental Algorithm, Artificial Intelligence, Operating System, Linear Algebra.

Skills

- Programming Skills: Python (Proficient), C/C++ (Proficient), SQL, Hadoop, AWS, Git, Matlab
- Data Analytic Skills: Machine Learning, NLP, Deep Learning, Web Scrawling, Data Visualization
- Packages: scikit-learn, genism, nltk, Theano, TensorFlow
- Languages: English, Chinese

Professional Experience

American International Group Inc. **New York, NY**
Research Assistant, Intern: License plate detector and heat map generator for damaged parts *Aug 2016-Dec 2016*
- Contribute to the license plate detection on low-resolution image dataset and heat map generation of damaged parts.
- Build an efficient license plate detector by fine-tuning a simple but efficient convolution neural network to generate saliency map and then use OpenCV to extract contour of license plate.
- Build an end-to-end Grad-CAM solution to extract information about damage part from a pre-trained model and an image.
- Implement these methods with Theano and Tensorflow and test them on both Linux and Windows system.
- The license plate detector achieves 15% top-5 error with 30s/image speed.

Academic Projects

Center for Data Science, NYU **New York, NY**
Machine Learning: Duplication Detection for health care information system *Feb 2016-May 2016*
- Build an end-to-end solution to detect duplicated record in health care information system.
- Use 32-bit rolling hash to compute the representative value of string quickly.
- Find sets of similar information entries by using K-means clustering and use elbow method to determine the number of sets.
- Create our own parallel filter algorithm to find possible duplication pairs in each set.
- Use T-SNE technique to visualize the feature vector and train a random forest model based on these vectors.

Machine Learning: Yelp Restaurant Rating Prediction *Feb 2016-May 2016*
- Build a model to predict future rating level of restaurant based on business attributes, previous ratings and Yelp reviews.
- Divide the rating into 3 different levels based on the distribution to make a balanced dataset.
- Use Google pre-trained word2vec model to represent Yelp reviews as 300-dimension vector.
- Select models and corresponding hyper-parameters by using cross-validation and AUC-scores.
- The best model is a logistic regression with L1 regularization and its average cross-validation AUC-scores is 0.86.

Anomaly Detection: Real Time Data Ingestion and Anomaly Detection for Particle Physics *Sep 2016-Dec 2016*
- Build an anomaly detector for detecting abnormal events from 100GB normal CERN particle collision event.
- Train a compressor and re-constructor for capturing insight feature of physics events by using deep neural auto-encoder.
- Evaluate auto-encoder by calculating reconstruction error (R2-score) and deep MLP auto-encoder has 0.95 R2-score.
- Detect abnormal event by checking whether reconstruction error from well-trained auto-encoder is greater than threshold.
- This work is supervised by Prof. Kyle Cranmer and presented in NIPS 2016 invited talk.

Big Data: Explore Relationship Between Citi-bike and weather *Feb 2016-May 2016*
- Extract insight relationship between Citi-bike and weather in New York City by using data in 2015.
- Filter and select different features from data concurrently by using Hadoop MapReduce in NYU HPC cluster.
- Visualize data by using different technologies such as google map API (gmaps).
- The conclusion agrees with our hypothesis – when the weather gets better, the duration of trips increases.