Zihao Wang

204 10th St, Apt 314, Jersey City, New Jersey, 07302

✉ zw1074@nyu.edu zw1074.github.io

☎: 551-(225)-9955

Education

- **New York University** — New York, USA
  Master of Data Science — *Sep. 2015 - Now*
  - **Graduation day**: Jun. 2017, **GPA:** 3.8/4.0
  - **Relevant Courses**
    Machine Learning, Deep Learning, Statistical Learning, Nature Language Processing, Inference and Graph Model, Time Theory, Logic
- **Nanjing University** — Nanjing, China
  Bachelor of Science, Computational Mathematics — *Sep. 2011–Jun. 2015*
  - **GPA**: 3.5/4.0

Skills and Interest

- **Programming Language:** C++/C, Python, Lua, Scheme, Mathematica, MATLAB
- **Technique:** Tensorflow, Theano, Torch7, CUDA C++/C, Hadoop, SQL
- **Interest:** Parallel Programming, Machine Learning, Deep Learning, Natural Language Processing, Computer Vision.

Experience

- **Research Assistant Intern (Deep Learning)** — AIG Inc, NYC USA
  Science Team — *Aug. 2016 - Dec. 2016*
  - Contribute to the whole automatic car damage appraisal project, especially for license plate detection and heat map generation of damage part.
  - Build an end-to-end solution for accelerating license plate detection. Also implement this solution by using Theano and OpenCV library.
  - Use a novel method to generate heat map. Design some experiments to test the effect. Help to make the method be compatible with both Windows and Linux system by using Theano and Tensorflow. Also make an end-to-end toolbox for efficiently using this method.

Projects

- **Efficient auto-encoder for physics particle collision event** — New York University, USA
  Teamwork: response for designing and implement — *Oct. 2016*
  - Use collision event data from CERN to produce an auto-encoder to compress data.
  - Compare three compressors: multilayer perceptron auto-encoder, convolutional auto-encoder and PCA by calculating the reconstruction error and applying reconstrcted data in real application. The best auto-encoder is multilayer perceptron, which has over 0.92 R2 score between the reconstructed and original data.
  - Add threshold RELU on the last layer to make the output sparse, which increase the R2 score from 0.92 to 0.95.
  - Use this technique to do anomaly detection and compare the mean square error between the normal one and the abnormal one. The multilayer perceptron auto-encoder is still the best one.
- **Duplication Detection** — New York University, USA
  Individual: response for designing and implement — *May. 2016*
  - Use the data from health care system to predict possible duplication of information. The whole pair set is around 10^{11} including 120,000 duplication pairs in ground-truth.
  - Construct an efficient parallel method to get a smaller set of candidate pair which is possibly duplicate. The amount of pair set is reduced from 10^{11} to $3700k$.
  - Generate a balanced training set by randomly selecting same number for two group. Extract feature vector for each pair by our business sense. Visualize the vectors by T-SNE technology. Compare different binary classification model. Random forest has the best performance
  - The smaller set of interesting pairs includes over 95% ground truth. And finally get around 94% accuracy with our fine tunning classifier.
- **Explore Relationship Between Citi bike and weather** — New York University, USA
  Teamwork: response for designing and implement — *May. 2016*
  - Use citi bike data and weather data in 2015 to find the relationship.
  - Create MapReduce functions to filter or edit dimension of data on Hadoop platform. In addition to test the relation between weather and citibike usage, also add the dimension such as age and gender to check if the results would vary for these groups.
  - Use data visualization technique to explore the correlation. The collusion is that temperature has very high correlation and the usage of citi bike is various depending on time, traffic, gender and age groups.
- **Yelp Restaurant Rating Prediction** — New York University, USA
  Teamwork: response for designing and implement — *Dec. 2015*
  - Use the data from Yelp Dateset Challenge to fit different models.
  - Create a new model by tagging words of each review as adjective then apply Google pre-trained word2vec model. Compare different multi-class classification algorithm and use micro-ROC (the average probability that the confident score of true sample is higher than the false sample) for evaluation.
  - By using word2vec model, it can improve the micro-AUC score by 50%. Logistic Regression has best performance (0.86 micro-AUC score).