Zihao Wang

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

Mail:zw1074@nyu.edu, Website:zw1074.github.io
Phone: 551-(225)-9955

Education

- **New York University (CIMS)** — 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, Theorem and Application in NLP, Generative Adversarial Network, Transfer Learning.

Experience

- **Research Assistant** — 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.
  - Help to build an end-to-end solution for accelerating license plate detection. Also implement this solution by using python 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. 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. The data is 3-D tensor while the index represent the location of the energy detector and the value represent the energy. The challenge is that this noisy data has 14400 dimension and we need to preserve the most relevant part and thus reduce the noise. The compress ratio is 32:14400.
  - Compare three compressors: Multilayer perceptron auto-encoder, convolutional auto-encoder and PCA. We evaluate them by two ways. One is calculating the reconstruction error and the other is apply reconstructed data in real application to see its performance.
  - Add threshold RELU on the last layer to make the output sparse. We find that multilayer perceptron is the best auto-encoder because it has over 0.95 AUC score between the reconstruction data and the original data. Also unlike PCA, it does not focus on the biggest value but indeed extract the hidden factor via our training process.
  - We use this technique to do anomaly detection. We compare the mean square error between the normal one and the abnormal one. And we find that multilayer perceptron performs best in this case.
- **Duplication Detection** — New York University, USA
  Teamwork: response for designing and implement — *May. 2016*
  - Use the data from health care system to predict possible duplication of information. The challenge is that the whole pair set is around 10^{11}, which is time consuming if we build the model on it. And it is also not a balanced dataset because in ground truth we only have around 120,000 pairs.
  - Construct an efficient parallel method to get a smaller set of interesting pairs, which we think that they are duplication. The processing time is 10 minutes using 8 workers on CPU. After this process, we reduce the pair set from 10^{11} to $3700k$. Then we generate a balanced training set by randomly select same number negative pairs (pairs that are not duplication) as the positive pairs (pairs that are duplication). Then we use a feature extractor to generate feature vector for each pair. Finally use random forest to make the prediction.
  - The smaller set of interesting pairs includes over 95% ground truth. And we finally get 99.4% 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.
  - I create multiple MapReduce functions to filter or edit dimension of data. Besides only testing the relation between weather and usage, we also added the dimension such as age and gender in order to check if the results would vary for these groups.
  - Use data visualization technique to explore the correlation. We find 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. The challenge of this dataset is that the business attribute of the restaurant is not enough for well prediction so that combining the review as the additional feature is necessary.
  - Create a new model by tagging words of each review as adjective then apply Google pre-trained word2vec model which can improve the accuracy by 50%. Also evaluate the model by using AUC of the micro-ROC curve, which is equal to the probability that the confident score of true sample is higher than the score of false sample. For our model, the AUC/probability is 0.86.