# Siyu Ren

Ph. D student
Shanghai Jiao Tong University, Shanghai, China
✉ : roy0702@sjtu.edu.cn
Language: CET-6 568

## Research Interests

- Language Models
- Model Compression, e.g., Knowledge Distillation, Weight Pruning
- NLP Applications

## Educational Background

- Shanghai Jiao Tong University, Computer Science (GPA: 3.7/4.0), PhD Candidate, 2019.09—now
- Tong Ji University, Computer Science (GPA: 4.7/5.0), Bachelor, 2015.09--2019.06

## Publications

### Model Compression for Efficiency:

The research purpose is to improve model efficiency in terms of storage, runtime memory, inference latency, and computation.

- **Siyu Ren**, Kenny Q. Zhu*: Pruning Pre-trained Language Models with Principled Importance and Self-regularization. Findings of ACL (CCF-A) 2023
    - Recast the importance criterion design in iterative pruning as an equality-constrained 0-1 ILP problem and derive a new principled importance score.
    - Propose a self-regularization training scheme to boost sparse model's generalization performance.
    - By exploiting resulting sparsity, the storage/inference can be reduced/accelerated by 8.9x and 2.x using CSR format and sparsity-aware runtime.

- **Siyu Ren**, Kenny Q. Zhu*: Leaner and Faster: Two-stage Model Compression for Lightweight Text-Image Retrieval. NAACL (CCF-B) 2022, 4085-4090
    - Propose a two-stage model compression framework tailored for light-weight text-image retrieval by exploiting abundant unpaired text/image data.
    - Open-sourced efficient image/text encoders with retrieval accuracy competitive with CLIP.
    - Open-sourced mobile (iOS and Android) applications.

### Model Compression for Knowledge Specialization:

The research purpose is to identify, extract and distil specialized knowledge from pre-trained language models to better understand their inner mechanisms.

- **Siyu Ren**, Kenny Q. Zhu*: Specializing Pre-trained Language Models for Better Relational Reasoning via Network Pruning. Findings of NAACL 2022 (CCF-B), 2195-2207
    - Proposes an end-to-end differentiable weight pruning framework for specializing general-purpose pre-trained language models into grounded commonsense relational models at non-trivial sparsity.
    - The resultant subnetworks exhibit higher generalization ability at scenarios requiring knowledge of single or multiple commonsense relations, e.g., knowledge base completion, commonsense reasoning.

### NLP Application:

The research purpose is to spot pain points in real-world applications and adapt cutting-edge NLP techniques to promote productivity.

- **Siyu Ren**, Kenny Q. Zhu*: Knowledge-Driven Distractor Generation for Cloze-style Multiple Choice Questions. AAAI 2021 (CCF-A): 4339-4347
    - Compile and open source a diverse and comprehensive benchmark dataset for training and evaluating distractor generation model.
    - Propose a configurable distractor generation framework for open-domain cloze-style MCQ, which requires no domain-specific vocabulary and jointly evaluates the plausibility and reliability of distractors.
    - Comprehensive experiments to evaluate and analyze various instantiations of our framework.

### Others:

- Qi Jia, Yizhu Liu, **Siyu Ren**, Kenny Q. Zhu*, Haifeng Tang: Multi-turn Response Selection using Dialogue Dependency Relations. EMNLP 2020 (CCF-B): 1911-1920

## Honors and Awards

- 2021-2022 GuangHua Scholarship
- 2017-2018 Undergraduate Excellent Student Second-Class Scholarship
- 2016-2017 Undergraduate Excellent Student Second-Class Scholarship