Yian Zhang

+86 136 7192 6669 | yian.zhang@nyu.edu | 1555 Century Ave, Pudong New District, Shanghai, China, 200122

EDUCATION

New York University Shanghai Expected graduate time: May 2021

Major：**Computer Science** | Minor: **Mathematics**

Cumulative GPA: 3.91/4.00 | Major GPA: 3.95/4.00

Selected Coursework: Machine Learning for Language Understanding (A), Artificial Intelligence (A), Parallel Computing (A) Natural Language Processing (A), Operating Systems (A), Basic Algorithms (A), Data Structures (A)

Scholarship: NYUSH 2019 Recognition Award | NYUAD 2019 Visiting Undergraduate Research Scholarship
NYUSH 2020 Recognition Award

PUBLICATIONS & PREPRINTS

[1] **Yian Zhang***, Alex Warstadt*, Haau-Sing Li, and Samuel R. Bowman. When Do You Need Billions of Words of Pretraining Data? *arXiv:2011.04946 preprint,* 2020.

[2] **Yian Zhang**. Latent Tree Learning with Ordered Neurons: What Parses Does It Produce? *The Workshop on Analyzing and interpreting neural networks for NLP (BlackboxNLP)*, 2020.

[3] Alex Warstadt, **Yian Zhang**, Haau-Sing Li, Haokun Liu, and Samuel R. Bowman. Learning Which Features Matter: RoBERTa Acquires a Preference for Linguistic Generalizations (Eventually). *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[4] Daniel Chin, **Yian Zhang**, Tianyu Zhang, Jake Zhao, and Gus Xia. Interactive Rainbow Score: A Visual-centered Multimodal Flute Tutoring System. *Proceedings of the International Conference on New Interfaces for Musical Expression(NIME)*,2020.

[5] **Yian Zhang**, Yinmiao Li, Daniel Chin, and Gus Xia. Adaptive Multimodal Music Learning via Interactive-haptic Instrument. *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*, 2019.

RESEARCH

Investigating the Amount of Pretraining Data Required to Learn Different NLU Skills

Research Assistant advised by Professor Sam Bowman ML^2, CILVR, NYU, May 2020 – Current

- We probe RoBERTas pretrained with 1M, 10M, 100M, 1B and 30B words using the classifier probing method, the information-theoretic probing method, unsupervised relative grammaticality judgement, and finetuning on NLU tasks
- The first three probing methods show that 90% of the attainable learning of syntax and semantics can be made with about 10M and 100M words of pretraining respectively, while commonsense learning requires much more data
- The last experiment shows that RoBERTa's applied NLU performance generally improves very little until the pretraining data volume exceeds 1B words, suggesting that some critical skills other than the basic syntactic and semantic knowledge we test cannot be learnt with <1B words, and commonsense reasoning is one such possibility
- Co-author a paper (as the **1st author**) which will be submitted to the conference of **ACL 2021**

Investigating the Impact of Pretraining on RoBERTa's Inductive Bias

Research Assistant advised by Professor Sam Bowman ML^2, CILVR, NYU, January 2020 – Current

- The project investigates how the amount of pretraining data affects RoBERTa's preference for linguistic generalizations over surface generalizations during finetuning on a downstream task
- We pretrain RoBERTa from scratch with 1M, 10M, 100M, 1B words and test them together with Facebook's RoBERTa$_{BASE}$ (pretrained on about 30B words) on MSGS, a dataset we create following the *poverty of the stimulus* design to probe the model's generalization preference
- We find that (1) the linguistic preference is closely related with pretraining data volume, and RoBERTa requires at least 30B words of pretraining to demonstrate a linguistic preference without external hints; (2) RoBERTa requires far more pretraining data to learn to prefer a linguistic generalization than to identify the relevant linguistic features
- Publish a paper (as the **2nd author)** at the conference of **EMNLP 2020**; a follow-up work that focuses on spoken language and structural bias will be submitted to the journal *Language*

Analysis of the Unsupervised Parsing Behavior of ON-LSTM

Course Project inspired by Professor Sam Bowman NYU, March 2020 – October 2020

- This project investigates the latent tree learning model ON-LSTM: how consistent is its unsupervised parsing behavior and how are its parses different from PTB gold parses?
- Reproduce ON-LSTM 5 times with different random seeds, compute the average F1 score between the parses of each pair of ON-LSTMs (self F1 score), and find that the model is reasonably consistent, with self F1= 65.7 on WSJ test and 82.1 on WSJ 10 compared to the random baseline's 24.8 and 40.7
- By computing the model's constituent-type-wise parsing accuracy and examining its parses, find that (1) it struggles with internal structures of complex noun phrases; (2) it has a tendency to branch right before verbs too early (~74% of the times); (3) both issues could potentially be attributed to the training task—unidirectional language modelling
- Publish a paper (as the **1st author**) at the **EMNLP 2020 workshop BlackboxNLP**

Interactive Multimodal Music Learning System

Advised by Professor Gus Xia Music X Lab, NYU Shanghai, April 2018 – Current

- The project aims to build an interactive learning environment that teaches flute playing by giving real-time haptic, audio, and visual feedbacks
- Design the innovative "Clutch" mechanism that allows instant adjustment of the haptic guidance level, and a dynamic learning algorithm that boosts the learning rate by 45.3% and shrinks the forgetting chance by 86%
- Build both the hardware and software of the system, and design and conduct user studies to test the learning effect
- Publish two papers as the **1st author and 2nd author** respectively at conferences **NIME 2019 and NIME 2020**

Haptodont: Haptic-based Dental Simulation

Advised by Professor Mohamad Eid AIM Lab, NYU Abu Dhabi, June 2019 – August 2019

- The goal of the project is to simulate an oral environment using virtual reality technologies, where dental students can interactively practice dental probing under haptic and visual guidance
- Design and implement the 3D recording mode, where the instructor's probing demonstration is captured in detail
- Design and implement the 3D playback mode, where real-time haptic and visual guidance are provided to the learner according to the difference between earner inputs and the recorded instructor demonstration
- Resolve the problems of abrupt force variation and oscillation, by developing a force transition smoothing function and a basic PID controller

ADDITIONAL INFORMATION

Languages: Native in Mandarin, Working proficiency in English

Programming: **Python, Pytorch,** C++, Java, Javascript

Hobbies: Piano, Kung Fu, Soccer, NBA