Yian Zhang

(650) 709-6573 | yianzh@stanford.edu | 450 Jane Stanford Way Stanford, CA 94305–2004 | yianzhang.github.io

EDUCATION

Stanford University *September 2021 – March 2023 (Expected)*

M.S. in Computer Science

New York University Shanghai *August 2017 - May 2021*

B.S. in Computer Science, Minor in Mathematics

Dean's Award (Top 1 of the CS department) | Cumulative GPA: 3.92/4.00 (*Summa cum laude*)

PUBLICATIONS

[1] When Do You Need Billions of Words of Pretraining Data? (**ACL 2021**) - **1st author** [URL]

[2] Latent Tree Learning with Ordered Neurons: What Parses Does It Produce? (**EMNLP 2020 workshop**) - **1st author** [URL]

[3] Learning Which Features Matter: RoBERTa Acquires a Preference for Linguistic Generalizations (Eventually). (**EMNLP 2020**) – **2nd author** [URL]

Check my Google Scholar Page for my other papers in Computer Music and Human Computer Interaction.

EXPERIENCES

Machine Learning Engineer Intern **ByteDance**

Intern at the Dialog System NLU Team *May 2021 - August 2021*

- Designed and built a multipurpose slot tagger from scratch and deployed it to aid customer intent recognition.
- Defined the slot taxonomy (covering 61% of all tokens) and crowdsourced >22K training and test examples.
- Used existing general NLP tools (SRL, POS, Dependency Parsing, etc.) for bootstrapping and dataset expansion.
- Trained a BERT-based slot tagging model with F1 = 92; Trained a random forest intent classifier taking predicted slots as inputs which outperformed or equaled the neural model in use on ¼ of all intent types.

Investigating the Impact of Pretraining Data Volume **ML^2, CILVR, NYU**

Research Assistant advised by Professor Sam Bowman *January 2020 - May 2021*

- Pretrained 24 RoBERTa models on 1M, 10M, 100M, and 1B words, and probed them using 6 styles of evaluation.
- Contributed to the Jiant framework (pulled) to support more model types and the Online Coding Paradigm.
- Found that most linguistic skills could be acquired with 0.3% of RoBERTa's original pretraining data, while linguistic bias and factual knowledge took much more data to learn.
- Published a paper at **ACL 2021 (1st author)**; Another at **EMNLP 2020 (2nd author).**

Interactive Multimodal Music Learning System **Music X Lab, NYU Shanghai**

Research Assistant Advised by Professor Gus Xia *April 2018 - May 2021*

- Built an interactive environment that taught flute playing by giving real-time haptic, audio, and visual feedbacks.
- Implemented the GUI, the motor controller, and the adaptive learning algorithm that boosted learning speed by 90%.
- Published a paper at NIME 2019 (1st author); Another at NIME 2020 (2nd author).

News Recommendation with Document Understanding **NYU Shanghai**

Capstone Project advised by Professor Wilson Tam *January 2021 - May 2021*

- Built a DSSM-fashion recommender system to predict click through rate (CTR) on the MIND dataset.
- Reproduced the NRMS model with Pytorch and improved its group AUC on MIND-small by 4.3% without increasing model size by using pretrained MLMs and multi-view learning.

Boring Blogs **NYU**

Course Project *January 2020 - May 2020*

- Using MongoDB, Express, React, and Node.js, built a blog platform where users could sign up and post articles.
- Used the TF-IDF algorithm to efficiently recommend to the users the articles they were least interested in.

Software Engineer Intern **AIM Lab, NYU Abu Dhabi**

Intern at the Haptodont Team *June 2019 - August 2019*

- Worked on building a VR application to help dental students learn *probing* from haptic and visual feedbacks.
- Used C++ and Chai3d to implement the 3D recording mode that recorded the instructor's demonstration and the practice mode that gave proper real-time guidance to the learner according to their behavior.
- Developed a force transition smoothing feature to mitigate abrupt force variations and oscillations.

SKILLS

Programming Languages:	Python, JavaScript, SQL, C, Bash Script, Java, C++, Latex, HTML, CSS
Frameworks & Tools:	Pytorch, HuggingFace, Jiant, Scikit-learn, TensorFlow, MongoDB, React, Express