Yinkai Wang

18650146958 | ywang88@gmu.edu
yinkaiw.github.io

EDUCATION

George Mason University Aug 2018 - Aug 2021
Computer Science Bachelor VSE bachelor Fairfax
- Honors/Awards: Dean's List(2018-2020)

Huaqiao University Aug 2017 - Aug 2022
Computer Science Bachelor VSE bachelor Xiamen

PUBLICATIONS

- Fahim Faisal, **Yinkai Wang**, Antonis Anastasopoulos. Dataset Geography: Mapping Language Data to Language Users. Research paper for ACL 2022 Theme Track. (In submission)
- **Yinkai Wang**, Antonis Anastasopoulos. On the Cross-Lingual Consistency of Named Entity Recognition Models. Student Abstract for AAAI-UC. (In submission)
- Yuanqi Du, **Yinkai Wang**, Fardina Alam, Yuanjie Lu, Xiaojie Guo, Liang Zhao, Amarda Shehu. Deep Latent-Variable Models for Controllable Molecule Generation. Research Paper for IEEE BIBM 2021. (In submission)
- **Yinkai Wang***, Kaiyi Guan*, Aowei Ding*, Yuanqi Du. Ensemble Machine Learning System for Student Academic Performance Prediction. Educational Data Mining (EDM) 2021, Workshop for Undergraduates (W4U).

RESEARCH EXPERIENCE

Multilingual Geospatial Language Expression Discovery
Research Assistant(Advisor: Prof. Antonios Anastasopoulos)
- Define cross-lingual consistency as the desirable property that two parallel sentences in two languages, which should in principle use the same-named entities are actually tagged with the same named entities.
- Focus on the notion of cross-lingual consistency for multilingual NER models, and show the importance of also using parallel data in multiple languages to evaluate NER models.
- Train by parallel multilingual datasets. Use mBert model+aligned to predict the consistency of multilingual datasets.

Diffusion Probabilistic Models for Protain Generation
Research Assistant(Advisor: Dr. Amarda Shehu)
- Use diffusion cloud probabilistic models to treat each amino acid as a point. Generated the whole protein as a point cloud. Use the heatmap to compare generated proteins and training proteins.
- Use short-range and long-range to evaluate the result. To improve the credibility of generated protein, implement a loss function that includes a short and long-range.

Predicting Minimum Inhibitory Concentration for Quaternary Ammonium Compounds w/ Machineing Learning
Research Assistant(Advisor: Dr. Amarda Shehu)
- Create three settings based on the 70 features of ~450 Quaternary Ammonium Compounds. Base on these settings, use machine learning models to predict four properties of molecules.
- Predict four Minimum Inhibitory Concentration values with ten regression machine learning models. Use feature selection to analyze the best features settings for every property.

Ensemble Machine Learning System for Student Academic Performance Prediction
Researcher
- Use an ensemble machine learning system to predict students' final grades with multiple students' performances.
- Due to the influence of COVID-19, the education of students faced a severe problem: It's much harder for teachers to help students and know students' learning conditions.
- This model consists of two components, the ensemble feature engineering module, and the ensemble prediction module. Extensive experiment results have shown the superiority of our model over other traditional machine learning models, both in stability, efficiency, and accuracy.

Deep Latent-Variable Models for Controllable Molecule Generation
Research Assistant(Advisor: Dr. Amarda Shehu)
- Proposed several deep latent-variable models to generate small molecules with desired molecular properties.
- The models operate under supervised, disentangled representation learning and leverage both graph representation learning to learn inherent constraints in the chemical space and inductive bias to connect chemical and biological space.
- The evaluations show that the models are a promising step in controllable molecule generation in support of cheminformatics, drug discovery, and other application settings.

PROFESSIONAL EXPERIENCE

Bytedance Apr 2021 - Jul 2021
Intern DevEco Beijing
- Focus on the base of the host app of android, which has a coupling relationship with most of the apps from bytedance.
- Create a mock setting environment implement to help QA test, greatly improve the efficiency of testing and publishing.
- Make a great and comfortable environment for all the developers who are developing microapp on bytedance.

Google Kaggle
Google Smartphone Decimeter Challenge
- Ranked top 20% on the Google Smartphone Decimeter Challenge w/ a group of five undergraduate students.
- Designed data clearning, preprocessing, data analysis, model selection, result evaluation and visualization pipeline.
- Mastered real-world data science challenge with machine learning pipeline and team collaboration.

SKILLS LIST

Programming skills: Python, Java, C, Kotlin, MIPS, Julia
Research Interests: Machine Learning, Deep Learning, AI for Science, Deep Graph Learning, Natural Language Processing