Yufei Weng

(619) 385-2819 • yufei.weng.application@gmail.com • github.com/YufeiWeng • linkedin.com/in/yufei-weng-965033224

EDUCATION

UC San Diego
Master of Science, Computer Science
Expected Mar 2026
Relevant Coursework: Natural Language Processing (In Progress)

UC San Diego
Bachelor of Science, Computer Science and Mathematics (Applied)
June 2024
Relevant Coursework: Reinforcement Learning (Top 10% of class), Deep Learning, Computer Vision, Linear Algebra, Algorithm

EXPERIENCE

BergLab: Machine Learning Research – Student Researcher *Apr 2024 – Present*

- Applied robust learning techniques using the WILDS framework, improving Out-of-Distribution generalization for historical tablet dating and **increasing the F1 score by 11%**.
- Designed a data splitting strategy that accounted for distribution shift, enhancing model robustness and interpretability in **real-world scenarios**.
- Developed and fine-tuned an ensemble model using Detectron2 for Arabic textline detection, achieving page-level error rates **two times lower** than Kraken's (an OCR system optimized for historical and non-Latin script material) and demonstrating **word retrieval accuracy within 5%** of manual ground truth on challenging multi-typeface datasets.
- Engineered an iterative learning approach for page detection from scanned documents, significantly improving Page-Level OCR performance in downstream tasks. The related project, in collaboration with Google, is currently under review for publication.

UC San Diego CSE100: Advanced Data Structures - Tutor *Sep 2023 – Jun 2024*

- Co-designed six programming assignments (PAs) and two projects
- Guided students in their debugging efforts by using various tools such as **gdb, valgrind, and gprof**
- Held over **160 lab hours** and helped over **240 students**

SKILLS

Programming Languages: Python • C++ • C • SQL
Libraries, Frameworks, and Tools: PyTorch • Transformer • Linux • Git

PUBLICATIONS

D. Chen, J. Tian, **Y. Weng**, T. Berg-Kirkpatrick, and J. Myerston. 2024. Classification of Paleographic Artifacts at Scale: Mitigating Confounds and Distribution Shift in Cuneiform Tablet Dating. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, pages 30–41, Association for Computational Linguistics.