DI WU

+1 (310)498-8193 ⋄ diwu@cs.ucla.edu ⋄ https://xiaowu0162.github.io/

EDUCATION

University of California, Los Angeles   Los Angeles, CA, USA
Ph.D. in Computer Science   09/2022 - present
UCLANLP Group (with Prof. Kai-Wei Chang)
- Research Interests: keyphrase generation, text generation evaluation, retrieval-enhanced language models

University of California, Los Angeles   Los Angeles, CA, USA
B.S. in Computer Science, Summa Cum Laude   08/2018 - 06/2022
- Overall GPA: **3.973/4.000**
- HSSEAS Dean's Honor List (8 times)

RESEARCH EXPERIENCE

UCLA NLP Lab   Los Angeles, CA, USA
Student Researcher, Supervisor: Prof. Kai-Wei Chang   02/2021 - present

- Research on neural keyphrase generation methods and their evaluation.

- Research on building robust and reliable evaluation metrics for text generation.

- Collaborate with Taboola. Using an internal dataset as the use case, experiment with keyphrase generation and adaptation methods and compare with the company's working pipeline.

The Ozcan Research Group, UCLA   Los Angeles, CA, USA
Student Research Assistant, Supervisor: Prof. Aydogan Ozcan   09/2019 - 06/2022

- Researched on virtual staining of skin tissues with an emphasis on Basal Cell Carcinoma. Designed methods to improve both single-image quality and temporal coherence of the predictions.

- Explored dataset engineering and data augmentation methods to mitigate class imbalance.

PUBLICATIONS

Wu, D., Ahmad, W. U., Zhang, D., Ramanathan, M. K., and Ma, X. (2024). Repoformer: Selective Retrieval for Repository-level Code Completion. *ICML 2024.*

Wu, D., Ahmad, W. U., and Chang, K. W. (2024). On Leveraging Encoder-only Pre-trained Language Models for Effective Keyphrase Generation. *LREC-COLING 2024.*

Li, Y., Pillar, N., Li, J., Liu, T., **Wu, D.**, Sun, S., Ma, G., de Haan, K., Huang, L., Zhang, Y. and Hamidi, S., 2024. Virtual histological staining of unlabeled autopsy tissue. Nature Communications, 15(1), p.1684.

Wu, D., Ahmad, W. U., and Chang, K. W. (2023). Rethinking Model Selection and Decoding for Keyphrase Generation with Pre-trained Sequence-to-Sequence Models. *EMNLP 2023.*

Kung, P., Yin F., **Wu, D.**, Chang, K. W., and Peng N. (2023). Active Instruction Tuning: Improving Cross-Task Generalization by Training on Prompt Sensitive Tasks. *EMNLP 2023.*

Wu, D., Yin, D, and Chang, K. W. (2023). KPEval: Towards Fine-grained Semantic-based Evaluation of Keyphrase Extraction and Generation Systems. *Preprint.*

Wu, D., Ahmad, W. U., and Chang, K. W. (2022). Pre-trained Language Models for Keyphrase Generation: A Thorough Empirical Study. *Preprint.*

Wu, D., Ahmad, W. U., Dev, S., and Chang, K. W. (2022). Representation Learning for Resource-Constrained Keyphrase Generation. *Findings of the ACL: EMNLP 2022.*

Li, J., Garfinkel, J., Zhang, X., **Wu, D.**, Zhang, Y., de Haan, K., Wang, H., Liu, T., Bai, B., Rivenson, Y., Rubinstein, G., Scumpia, P., and Ozcan, A. (2021). Biopsy-free in vivo virtual histology of skin using deep learning. *Light: Science & Applications, 10*(1), 1-22.

INTERNSHIP EXPERIENCE

AWS AI   New York, NY, USA
Applied Scientist Intern; Mentor: Wasi Ahmad, Dejiao Zhang   06/2023 - 09/2023

- Researched on improving retrieval-augmented code language models for repository-level code completion.
- Formulated the task of selective retrieval-augmented infilling. Designed approaches from the perspective of in-repository code retrievers and the code generator models.
- The designed model achieves no performance loss with only 10% of retrieval budget.

Microsoft Research Asia   Beijing, China
Research Intern; Mentor: Ning Shang   04/2021 - 09/2021

- Participated in the research project "Sparse Analysis". Compared over 10 model compression methods including pruning, quantization, and knowledge distillation on MobileNetV2 and Transformers. For pruning, the analysis involved criterion, sparsity scheduling, and learning scheduling.
- Participated in the open source AutoML project NNI. Contributed over 5000 lines of code including a hyperparameter optimization benchmark, a Transformer pruner, and three examples.

NewsBreak   Beijing, China
Natural Language Processing Intern   07/2020 - 10/2020

- Worked on a hierarchical multi-label classification problem with 268 categories. Improved the f1-score by 49% with multiple statistical and deep learning methods.
- Improved the performance of model pre-training, fine-tuning, and online serving pipelines.

TEACHING

Teaching Assistant

- UCLA CS 33, Introduction to Computer Organization, Fall 2023, with Prof. Glenn Reinman.
- UCLA CS 33, Introduction to Computer Organization, Spring 2024, with Prof. Glenn Reinman.

SERVICES

- Reviewer: ACL 2023, EMNLP 2023, AAAI 2023/2024, JAIR, ACL Rolling Review.