Manli Shu

Palo Alto, CA 94301 • Google Scholar • azure.shu1997@gmail.com • (240) 714-2447

Education

University of Maryland, College Park *09/2019 – 12/2023*
Ph.D. in Computer Science, Department of Computer Science *College Park, MD*
Dissertation: Towards Trustworthy Machine Learning Systems

University of Science and Technology of China *09/2015 – 07/2019*
B.Eng. in Information Security, School of Information Science & Technology *Hefei, China*

National Tsing Hua University *02/2017 – 06/2017*
Exchange student, Department of Electrical Engineering *Hsinchu, Taiwan*

Work Experience

Salesforce, Applied Scientist, AI Research *01/2024 - Present*

- **Research Project:** Instruction fine-tuning of large multimodal models (BLIP-3). *Palo Alto, CA*
  - Project co-lead: Leading the post-training and open-source efforts of Salesforce's latest multimodal LLMs.
  - Built the post-training codebase and evaluation pipeline for BLIP-3 that supports large-scale interleaved image-text and video input.
  - Developed a dynamic high-resolution image encoding algorithm that improved average OCR performance by over 10%.
  - Responsible for open-sourcing the model and training code. Merged the BLIP-3 code to a third-party evaluation codebase.
- **Applied Project:** LLM-based code generation for Tableau data analysis.
  - Generating data visualization code from user query for Tableau Cloud.
  - Coordinated with the data annotation team and curated the in-house training data for fine-tuning Tableau LLM .
  - Designed and conducted human evaluation and error analysis on real user feedback.

Google, Research Intern *06/2023 - 11/2023*

- **Intern project:** Diffusion models for face anti-spoofing. *Mountain View, CA*
  - Designed a solution for face anti-spoofing using generative text-to-image diffusion models.
  - Validated the proposal with its zero-shot performance comparable to supervised baselines, with the additional benefit of providing explainable visualizations from the diffusion model.

Salesforce, Research Intern *06/2022 - 11/2022*

- **Intern project:** : Improving 3D vision transformers across object scales. *Palo Alto, CA*
  - Proposed a novel attention mechanism to improve the precision of 3D object detectors on small objects.
  - Improved the state-of-the-art transformer-based 3D detector on indoor benchmarks with an overall increase of 2.0% in mAP and 3.5% relative improvements on small objects.

Nvidia, Research Intern *01/2022 - 05/2022*

- **Intern project:** Prompt tuning for zero-shot generalization in vision-language models. *Remote, U.S.*
  - Developed a test-time self-supervised optimization objective for prompt tuning without downstream data or annotations.
  - Increased the out-of-distribution accuracy of a pre-trained vision-language model by 5.6%.

Research Experience

UMD Center for Machine Learning, Graduate Assistant *08/2019 - 12/2023*
Advisor: Tom Goldstein *College Park, MD*

- **Research Project:** Data safety research on the instruction tuning for LLMs.
- **Research Project:** Explainability for vision and language models.

Selected Publications

[1] L. Xue*, **M. Shu***, A. Awadalla, J. Wang, A. Yan, S. Purushwalkam, H. Zhou, V. Prabhu, Y. Dai, M. S Ryoo, S. Kendre, J. Zhang, C. Qin, S. Zhang, C. Chen, N. Yu, J. Tan, T. Awalgaonkar, S. Heinecke, H. Wang, Y. Choi, L. Schmidt, Z. Chen, S. Savarese, J. C. Niebles, C. Xiong, R. Xu. xGen-MM (BLIP-3): A Family of Open Large Multimodal Models. In *EVAL-FoMo @ ECCV*, 2024.

[2] A. Awadalla, L. Xue, O. Lo, **M. Shu**, H. Lee, E. Kumar Guha, M. Jordan, S. Shen, M. Awadalla, S. Savarese, C. Xiong, R. Xu, Y. Choi, L. Schmidt. MINT-1T: Scaling Open-Source Multimodal Data by 10x: A Multimodal Dataset with One Trillion Tokens. In *NeurIPS*, 2024.

[3] **M. Shu**, J. Wang, C. Zhu, J. Geiping, C. Xiao, T. Goldstein. On the Exploitability of Instruction Tuning. In *NeurIPS*, 2023.

[4] **M. Shu**, W. Nie, D. Huang, Z. Yu, T. Goldstein, A. Anandkumar, C. Xiao. Test-Time Prompt Tuning for Zero-Shot Generalization in Vision-Language Models. In *NeurIPS*, 2022.

[5] Y. Xu, J. Yao, **M. Shu**, Y. Sun, Z. Wu, N. Yu, T. Goldstein, F. Huang. Shadowcast: Stealthy Data Poisoning Attacks against Vision-Language Models. In *NeurIPS*), 2024.

[6] J. Geiping, A. Stein, **M. Shu**, K. Saifullah, Y. Wen, T. Goldstein. Coercing LLMs to Do and Reveal (Almost) Anything. In *SeT LLM @ ICLR*, 2024.

[7] N. Jain, K. Saifullah, Y. Wen, J. Kirchenbauer, **M. Shu**, A. Saha, M. Goldblum, J. Geiping, T. Goldstein. Bring Your Own Data! Self-Supervised Evaluation of Large Language Models. In *COLM*, 2024

[8] J. Kirchenbauer, J. Geiping, Y. Wen, **M. Shu**, K. Saifullah, K. Kong, K. Fernando, A. Saha, M. Goldblum, T. Goldstein. On the Reliability of Watermarks for Large Language Models. In *ICLR*, 2024

[9] Goldblum, M., Souri, H., Ni, R., **Shu, M.**, Prabhu, V., Somepalli, G., ... Goldstein, T. Battle of the backbones: A large-scale comparison of pretrained models across computer vision tasks. In *NeurIPS*, 2023.

[10] A. Ghiasi, H. Kazemi, E. Borgnia, S. Reich, **M. Shu**, M. Goldblum, A. G. Wilson, T. Goldstein. What Do Vision Transformers Learn? A Visual Exploration. (*Preprint. Under review.*)

[11] R. Levin*, **M. Shu***, E. Borgnia*, F. Huang, M. Goldblum, T. Goldstein. Where do models go wrong? Parameter-space saliency maps for explainability. In *NeurIPS*, 2022.

[12] R. Ni*, **M. Shu***, H. Souri, M. Goldblum, T. Goldstein. The Close Relationship between Contrastive Learning and Meta Learning. In *ICLR*, 2022.

[13] **M. Shu**, Z. Wu, M. Goldblum, T. Goldstein. Encoding Robustness to Image Style via Adversarial Feature Perturbations. In In *NeurIPS*, 2021.

[14] **M. Shu**, Y. Shen, M. Lin, T. Goldstein. Adversarial Differentiable Data Augmentation for Autonomous Systems In *ICRA*, 2021.

[15] **M. Shu**, L. Xue, N. Yu, R. Martín-Martín, C. Xiong, T. Goldstein, J. C. Niebles, R. Xu. Model-Agnostic Hierarchical Attention for 3D Object Detection. In *International Conferences on Robotics and Automation (ICRA), 2024.*

Technical Skills

- **Coding and Tools**: Python (PyTorch, TensorFlow, JAX), Git, Docker, Kubernetes, Gradio.
- **Machine Learning (AI/ML)**: Multimodal foundation models, Large language models, AI Safety and alignment,