PUNING YANG

(+86)-183-2202-8055 Shenzhen, China — puning97yang@gmail.com

EDUCATION

TMLR Group, Department of CS, Hong Kong Baptist University Jul. 24 - Present
Research Assistant Shenzhen, CN
Advisor: Ass. Prof. Bo Han

University of Chinese Academy of Sciences & Institute of Automation, CAS Sep. 21 - Jun. 24
Master in Electronic and Information Engineering (GPA 3.6/4.0) Beijing, CN
Advisor: Prof. Ran He (IEEE / IAPR Fellow)

Nankai University Sep. 2016 - Jun. 2020
Bachelor of Intelligence Science and Technology (GPA 75.6/100) Tianjin, CN

RESEARCH EXPERIENCE

Large Language Model Unlearning 2024.04-Present

***Puning Yang**, Qizhou Wang, Zhuo Huang, Tongliang Liu, Chengqi Zhang, Bo Han. Exploring Criteria of Loss Reweighting to Enhance LLM Unlearning.* (ICML submission, 2025)

- We first systematically investigate the role of loss reweighting in LLM unlearning by identifying two distinct goals: Saturation and Importance. We find that saturation-based reweighting is generally more effective than importance-based strategies, and their combination provides additional improvements.
- Our investigation into specific reweighting operations revealed that the smoothness and granularity of weight distributions significantly influence unlearning performance.
- We proposed SatImp, a simple yet effective reweighting method that integrates the strengths of both saturation and importance, while carefully choosing specific reweighting operations.

*Qizhou Wang, Bo Han, **Puning Yang**, Jianing Zhu, Tongliang Liu, Masashi Sugiyama. Unlearning with Control: Assessing Real-world Utility for Large Language Model Unlearning.* (ICLR 2025)

- Regarding that existing metrics in LLM unlearning are not comprehensive, we propose a more robust metric: Extraction Strength (ES).
- We propose a simple yet effective model mixing method which achieves significantly better trade-offs between unlearning and retention.

Out-of-Distribution Detection 2022.10-2023.11

***Puning Yang**, Jian Liang, Jie Cao, Ran He. AUTO: Adaptive Outlier Optimization for Online Test-Time OOD Detection.* (Arxiv 2303.12267 Citation: 19)

- We propose a more practical setting for OOD detection, which considers an online test data stream and more complex components of test OOD data.
- We aim to optimize the OOD detector while making predictions during testing. The proposed framework, namely Adaptive oUTlier Optimization (AUTO), achieved superior performance in OOD detection and ID classification tasks. AUTO reveals the huge potential of utilizing test data to solve the OOD detection problem.

Face Forgery Detection 2021.09-2023.05

*Lixia Ma (advisor), **Puning Yang**, Yuting Xu, Ziming Yang, Peipei Li, Huaibo Huang. Deep learning technology for face forgery detection: A survey.* (Neurocomputing)

***Puning Yang**, Huaibo Huang, Zhiyong Wang, Aijing Yu, Ran He. Confidence-Calibrated Face Image Forgery Detection with Contrastive Representation Distillation.* (ACCV 2022)

- Based on the adverse idea, we propose to use neural networks to learn the universal features of different face generation paradigms (face swapping or face reenactment). Features are fused into the same network by knowledge distillation.
- Considering that both single-deepfake detectors are overconfident, we propose calibrating the knowledge distillation process with label smoothing and dynamic confidence weights. Results show that our proposed method achieves SOTA performance for cross-dataset face forgery detection.

ACADEMIC SERVICES

Reviewer NIPS(2022, 2023, 2024, 2025), ICML(2023, 2024, 2025), ICLR(2023, 2024, 2025), CVPR(2023, 2024, 2025), ICCV(2023, 2025), IEEE T-IP, IEEE T-CSVT