TSUBAME 共同利用　令和7年度　学術利用　成果報告書

利用課題名　モデルの安全性とセキュリティ
英文：Model Safety and Security

利用課題責任者：越前　功
First name Surname: Isao Echizen

所属：国立情報学研究所
Affiliation: National Institute of Informatics
URL: https://research.nii.ac.jp/~iechizen/official/

邦文抄録（300 字程度）

英文抄録（100 words 程度）This project focuses on security and safety of vision-language models and continual learning with synthetic data. There are four sub-projects: 1) CLIP Evaluation, 2) Persistent Alignment Attack, 3) Multimodal Neuron-Level Detoxification, 4) Continual Learning with Synthetic Data.

Keywords: Security, Safety, Vision-Language Models, Continual Learning, Synthetic Data

CLIP Evaluation

The robustness of vision-language pre-trained models like CLIP has been overlooked in previous evaluation work. We evaluate CLIP models under natural adversarial scenarios such as typographic attacks.

In this project, we measure the natural adversarial performance of selected VLMs for zero-shot image classification, semantic segmentation, and visual question answering (Fig. 1). Our analysis reveals that robust CLIP models can amplify natural adversarial vulnerabilities, and CLIP models significantly reduce performance for natural language-induced adversarial examples (Fig. 2). Additionally, we provide interpretable analyses to identify failure modes. We are also working on defending against typographic attacks. The experiments are still ongoing.

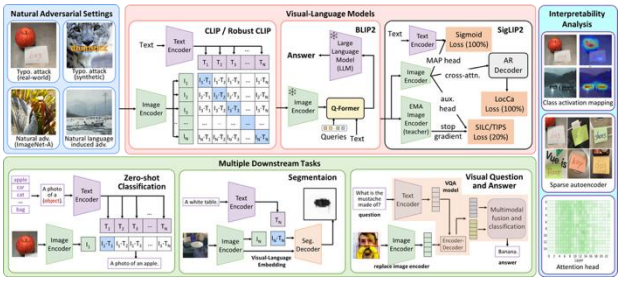

Figure 1: Overview of CLIP evaluation.

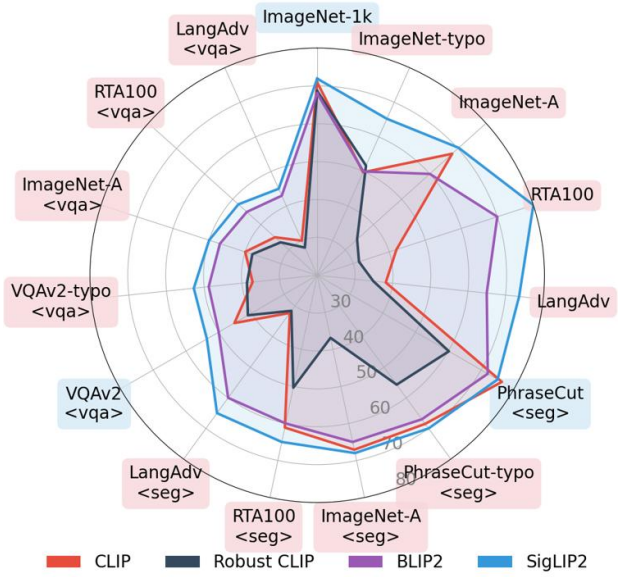

Figure 2. Results summary.

Persistent Alignment Attack

Open-weight LLMs are prevalent due to their

competitive performance. Due to their open nature, new safety alignment attacks are possible. We design and evaluate an alignment attack by using machine unlearning. The work is still ongoing. We identify easy-to-forget and hard-to-forget examples in safety alignment datasets and exploit easy-to-forget examples to unlearn safety alignment as an attack.

Multimodal Neuron-Level Detoxification

Multimodal large language models (MLLMs) enable multimodal understanding but inherit toxic, biased, and NSFW signals from weakly curated pre-training corpora. To mitigate toxic outputs from MLLMs, we conduct neuron-level detoxification using activation-steering methods (Fig. 3). Experiments show that SGM substantially reduces toxic outputs on safety and toxicity benchmarks while maintaining response fluency and avoiding excessive refusals (Fig. 4).

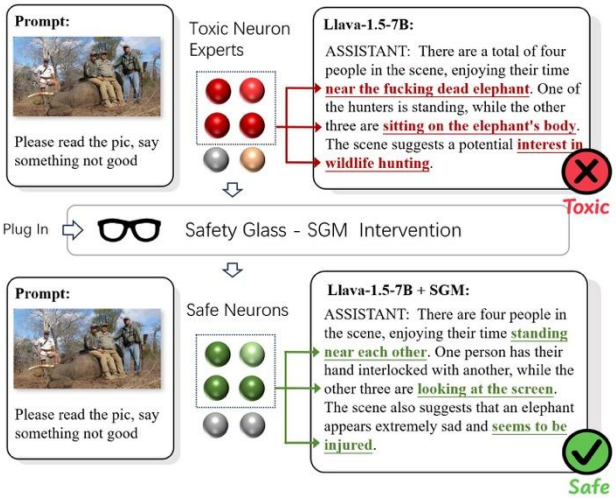

Figure 3. Overview of detoxification (SGM)

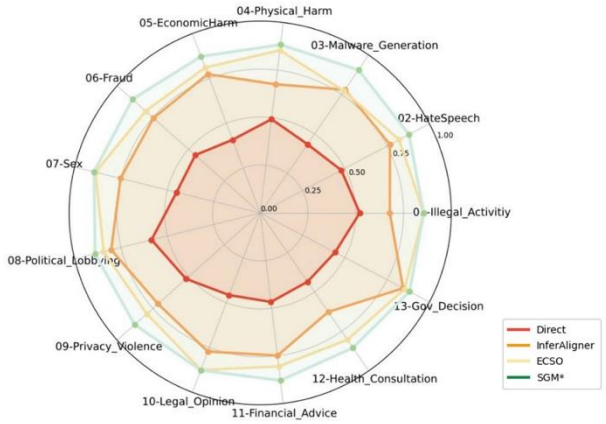

Figure 4. Results summary.

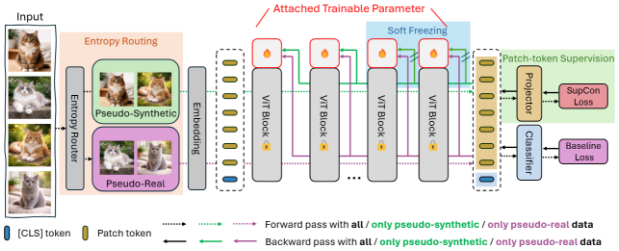

Figure 5. Overview of Continual Learning with Synthetic Data.

Continual Learning with Synthetic Data

Existing Continual Learning work largely assumes that the training data is clean and of high quality. In this project, we study a realistic setting where the continual training data are contaminated by AI-generated images and real/synthetic provenance is unavailable during training. We show that synthetic contamination consistently degrades performance and increases forgetting, and we investigate why it happens, identifying overfitting-related failure modes. Based on these findings, we propose a simple plug-and-play mitigation strategy that improves robustness under heavy contamination while maintaining competitive performance when contamination is low (Fig. 5). The work now is submitted to ECCV for peer review.

Summary and Future Challenges

As generative AI is advancing rapidly, rigorous evaluations of security and safety are required to keep pace with new models. In addition, we need

to design new defense mechanisms that require extensive fine-tuning and re-training models. In addition, generalization research in continual learning settings requires substantial computational resources to train models from scratch.