School of Computing, Engineering, and the Built Environment
Edinburgh Napier University

PHD STUDENT PROJECT

Application instructions:
Detailed instructions are available at :
https://www.napier.ac.uk/research-and-innovation/doctoral-college/how-to-apply

Prospective candidates are encouraged to contact the Director of Studies (see details below) to discuss the project and their suitability for it.

Project details

Supervisory Team:

- DIRECTOR OF STUDY: Dr Zhiyuan Tan (Email: z.tan@napier.ac.uk)
- 2ND SUPERVISOR: Dr Yanchao Yu

Subject Group: Cyber Security and System Engineering

Research Areas: Cyber Security, Artificial Intelligence, Machine Learning

Project Title: Machine unlearning for large language models

Project description:

ChatGPT has popularized Large Language Models (LLMs), which are now at the forefront of today's AI boom. The LLMs are remarkable tools that can understand, generate, and interact with human language in a natural way. However, these models often generate problematic responses due to improper training on harmful text from the internet. They also memorize and release copyright-protected content, and frequently deliver factually incorrect responses that can mislead users.

Therefore, it is crucial to generate safe outputs that align with human values and policy regulation, which is a significant challenge for LLM practitioners. Recent research [1-3] has shown that it is possible to remove undesirable behaviours in machine learning models without costly re-training. This PhD project aims to study and develop the ability to perform unlearning, which is the process of forgetting undesirable behaviours in LLMs, such as (1) removing harmful responses, (2)

erasing copyright-protected content as requested, and (3) eliminating hallucinations from LLMs.

You can find more information about Machine Unlearning at [5].

References:

[1] Cao, Y., & Yang, J. (2015, May). Towards making systems forget with machine unlearning. In 2015 IEEE Symposium on Security and Privacy (pp. 463-480). IEEE.

[2] Ginart, A., Guan, M. Y., Valiant, G., & Zou, J. (2019). Making ai forget you: Data deletion in machine learning. arXiv preprint arXiv:1907.05012.

[3] Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., ... & Papernot, N. (2019). Machine unlearning. arXiv preprint arXiv:1912.03817.

[4] Hu, H., Salcic, Z., Dobbie, G., & Zhang, X. (2021). Membership Inference Attacks on Machine Learning: A Survey. arXiv preprint arXiv:2103.07853.

[5] https://github.com/jjbrophy47/machine_unlearning

[6] Papernot, N., McDaniel, P., Sinha, A., & Wellman, M. P. (2018, April). Sok: Security and privacy in machine learning. In 2018 IEEE European Symposium on Security and Privacy (EuroS&P) (pp. 399-414). IEEE.

[7] Yao, Y., Xu, X., Liu, Y.: Large language model unlearning (2023) http://export.arxiv.org/abs/2310.10683

Candidate characteristics

Education:

A second class honour degree or equivalent qualification in Computer Science or Data Science with a good fundamental knowledge of machine learning.

Subject knowledge:

- Machine learning
- Natural language processing
- Python Programming

Essential attributes:

- Experience of fundamental machine learning
- Competent in programming and critical analysis
- Knowledge of security and privacy of machine learning
- Good written and oral communication skills
- Strong motivation, with evidence of independent research skills relevant to the project
- Good time management

Desirable attributes:

- Programming experience in Python and Machine Learning frameworks (e.g., TensorFlow or Keras)
- Good knowledge of deep learning, natural language processing, etc.
- Experience in Generative AI