Contribution ID: **2** Type: **Poster**

Comparative Study of Machine Unlearning Techniques for Computer Vision and NLP Models

Machine unlearning is an emerging field in machine learning that focuses on efficiently removing the influence of specific data from a trained model. This capability is critical in scenarios requiring compliance with data privacy regulations or when erroneous data needs to be removed without retraining from scratch. In this study, I explore the importance of machine unlearning as a way to enhance privacy simultaneously not affecting the efficiency of machine learning models. Using the CIFAR- 10 and CIFAR-100 dataset, I implemented various unlearning methods like retraining on the retained set, instruction fine tuning a LLM model to forget biased sentences and distillation techniques. These methods allowed the models to forget specific contexts while not comprising on the model accuracy. My implementations yielded promising results in terms of unlearning effectiveness and I have used various unlearning metrics to compare with my implementation and the baseline performance. The outcomes demonstrate the potential these methods have to balance between privacy and model accuracy effectively.

Author: ROY, Mahule (National Institute of Technology Karnataka Surathkal)

Presenter: ROY, Mahule (National Institute of Technology Karnataka Surathkal)

Session Classification: Poster session