Doctoral School of Information and Biomedical Technologies
Polish Academy of Sciences (TIB PAN)

SUBJECT:

Data ownership and privacy meet generative neural networks

SUPERVISOR:

dr hab. inż. Paweł Morawiecki
pawel.morawiecki@gmail.com
Instytut Podstaw Informatyki PAN, ul. Jana Kazimierza 5, Warszawa

DESCRIPTION:

The project lies at the intersection of privacy, security and machine learning.
In recent years, there have been rapid advances in generative modeling techniques within the field of deep learning. Among these, generative diffusion models, particularly those utilizing the Stable Diffusion framework, have gained prominence due to their capability to generate high-quality, diverse, and intricate samples. These models hold considerable potential for numerous applications, such as data augmentation, art creation, and design optimization. However, as these models become more widely adopted, addressing privacy and data ownership concerns becomes essential.

One critical issue that arises in this context is determining whether a specific data point was used during the training process of a model. Extracting this information from a model can be crucial in cases where copyrighted or sensitive data are used without permission, leading to potential legal issues. The importance of these matters is reflected in the European Union General Data Protection Regulation (GDPR), particularly Article 17 often referred to as the "right to be forgotten".

In this project, we want to investigate whether it is possible to infer meaningful information on training set for big, real-life generative neural networks. We also are interested in unlearning (forgetting) a given image to ensure that the neural network is no longer able to generate a very similar image or its style. Such a precise forgetting, if successful, would allow a user to be forgotten/withdrawn from the service without the need to retrain the whole network.

The project will be realized within NCN OPUS grant (No: 2023/49/B/ST6/02580). **The scholarship is 5000 PLN plus 3500 PLN (scholarship from TiB Doctoral School).**

Desired candidate's skills: practice in computer programming (knowledge of programming in Python and experience in using the deep learning frameworks such as PyTorch, fluency at English

BIBLIOGRAPHY:

[1] Jan Dubinski, Antoni Kowalczuk, Stanislaw Pawlak, Przemyslaw Rokita, Tomasz Trzcinski, Pawel Morawiecki: Towards More Realistic Membership Inference Attacks on Large Diffusion Models. WACV 2024: 4848-4857

[2] Matthew Jagielski, Om Thakkar, Florian Tramèr, Daphne Ippolito, Katherine Lee, Nicholas Carlini, Eric Wallace, Shuang Song, Abhradeep Guha Thakurta, Nicolas Papernot, Chiyuan Zhang: Measuring Forgetting of Memorized Training Examples. ICLR 2023