Data Ownership and Privacy Meet Generative Neural Networks

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. Recently, Getty Images filed a lawsuit against Stability AI, accusing it of *unlawfully copying and processing millions of copyright-protected images*.

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".

The current best generative models are text-conditioned, namely, a user gives a model a prompt, e.g. "a painting of a spaceship in a style of Van Gogh" and a model generates the image. These models can easily imitate the style of a given artist, but they are also flexible enough to modify it or put an image in an unusual context (see Fig. 1).

Figure 1: Modern large generative models very rarely blindly copy the images from a training set. Instead, they flexibly extract concepts such as style, tone, or texture to generate new images. An original piece of art by Zdzisław Beksiński (left) vs an image generated by the generative model Midjourney with a prompt ‘Beksinski’s vision of love’ (right). The generated image preserves the characteristic features of Beksiński’s art, namely particular textures and omnipresent decay.

This leads to a situation where it is extremely difficult to infer whether a given sample was used in training or not. Therefore, data ownership is hard to establish and privacy concerns start appearing.

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.