Privacy in Deep Learning: Deleting Training Data from Neural Networks on Demand

Deep neural networks (DNNs) can achieve exceptional performance when they are trained with large volumes of data. Considering the scale of training data retrieved online, essentially there might be personal data collected from individuals without explicit consent. This raises significant data protection and privacy concerns on current machine learning algorithms, in the face of adversaries that are capable of extracting training data or implementing successful membership inference attacks. Current regulations on this established in the European Union, such as the General Data Protection Regulation (GDPR), enforces the requirement that individuals can request to have their data to be deleted (i.e., hold the *right to be forgotten*). Alongside removing such data examples from the training set, instead of re-training the whole model from scratch, we consider developing efficient, privacy-enforcing algorithms to adapt pre-trained models via a so-called *machine unlearning* approach. In this project we will focus on this privacy aspect of deep learning, and develop algorithms to tackle this problem in different application settings.

Goals & Tasks

- Review of state-of-the-art on data deletion methods from pre-trained DNNs.
- Simulating and benchmarking existing methods on a deep learning application.
- Implementation and experiments with novel unlearning algorithms.

Contact

Ozan Özdenizci
ozan.ozdenizci@igi.tugraz.at

Qualifications

- Interest in ML security and privacy.
- Experience with the Python based deep learning framework PyTorch is beneficial.
- Registered to one of the following:

  □ **Bachelor Thesis**
  ✓ **Seminar Project**
  ✓ **Master Thesis**