Machine Unlearning for Digital Pathology

Master Thesis Proposal
Supervisors: Dr. Reza Nasirigerdeh and Prof. Dr. Peter Schüffler

Machine unlearning [1,2] is the process of eliminating the effect of a given set of samples, known as the **forget set**, from a pretrained model. The aim of machine unlearning is to fulfill the **"right to be forgotten"**, specified in Article 17 of General Data Protection Regulation (**GDPR**) [3], which grants the individuals such as patients the right to ask data holders including hospitals and medical centers to have their data "forgotten". "Data forgetfulness" [4] in this regard implies that the data holders are obligated to not only remove the data of the given individuals from their storage systems but also unlearn the contribution of that data from any trained model.

The most naive approach to machine unlearning is "**exact unlearning**", where the model is retrained from scratch (with randomly initialized weights) on the dataset excluding the forget set. However, this approach is computationally expensive, especially for large (foundation) models and/or frequent unlearning requests. "**Approximate unlearning**" [2] addresses this challenge by performing the unlearning process in a more computationally efficient fashion.

To evaluate the performance of approximate unlearning algorithms, the model obtained from the exact unlearning is employed as the gold standard. Given that, more similarity between the models from a given approximate unlearning algorithm and exact unlearning implies better unlearning. To measure the similarity, we typically compare the accuracy or AURoC of the models separately on the **forget set**, **retain set** (whole dataset excluding the forget set), and the **test set**.

Previous studies including **SalUn** [5] and **DEL** [6] mainly focus on proposing new approximate unlearning algorithms, evaluating the algorithms on non-medical datasets such as CIFAR-10 [7] and ImageNet [8]. To the best of our knowledge, the application of machine unlearning on medical data, especially on **histopathology images** is still **underexplored**.

This project aims to investigate the performance of the existing **approximate unlearning** algorithms on **pathology images**, focusing on the patch-level and

slide-level **classification** tasks. Unlike the datasets used in many previous studies, where the samples are considered to be uncorrelated (i.e. belonging to separate individuals/entities), the samples, e.g. patches, in pathology images might belong to the same patient. This characteristic of pathology data makes the unlearning process more challenging. This is because even if a few patients ask for data forgetfulness, the contribution of many patches corresponding to their whole slide images must be unlearned from the model, which might lead to "**catastrophic unlearning**", in which the unlearned model might not be generalizable anymore.

This project is based on **foundation models** including the Vision Transformer (**ViT**) architecture [9] (**ViT_small** and **ViT_base** versions). The steps of the project are as follows:

1. **CRC** [10], **MHIST** [11], and/or **CCRCC** [12] are employed as **datasets**, where each dataset is divided into three subsets: the **retain** set, **forget** set, and **test** set. The images/patches of a given patient must belong only to one of the subsets.
2. The models are trained/fine-tuned using **self-supervised frameworks** such as **DINOV2** [13] and/or **MAE** [14] on the train set (retain set + forget set).
3. **Exact unlearning, full fine-tuning**, **SalUn**, and **DEL** are employed as the **unlearning algorithms**, taking the pretrained model, retain set, forget set, and algorithm-specific hyper-parameters as arguments, and aim to unlearn the contribution of the forget samples from the model, while still keeping the contribution of the retain samples.
4. The unlearned models from approximate unlearning (i.e. full fine-tuning, SalUn, and DEL) are compared with the model from exact unlearning on the retain set, forget set, and test set. The smaller difference between the accuracy/AURoC values from a given approximate unlearning algorithm and exact unlearning on the retain and forget sets implies better unlearning, and on the test set means better generalizability.
5. The experiments are repeated with **different unlearning ratios (**i.e. different sizes of the forget set e.g. 10%, 20%, and 30% of all patients) to investigate the **catastrophic unlearning** phenomenon in the context of pathology image data.

References

1. Cao, Y. & Yang, J. Towards making systems forget with machine unlearning. 2015 IEEE symposium on security and privacy, 463–480 (IEEE, 2015).
2. Bourtoule, L. et al. Machine unlearning. 2021 IEEE Symposium on Security and Privacy (SP), 141–159 (IEEE, 2021).
3. Voigt, P. & Von dem Bussche, A. The eu general data protection regulation (gdpr). A Practical Guide, 1st Ed., Cham: Springer International Publishing 10, 10–5555 (2017).
4. Nasirigerdeh et al. Machine Unlearning for Medical Imaging, arXiv preprint, https://arxiv.org/abs/2407.07539
5. Fan, C. et al. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. The Twelfth International Conference on Learning Representations (2023).
6. Torkzadehmahani et al. Improved Localized Machine Unlearning Through the Lens of Memorization, arXiv preprint, https://arxiv.org/pdf/2412.02432
7. Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.
8. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pp. 248–255. Ieee, 2009.
9. Dosovitskiy, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020).
10. Oliveira, Sara P., et al. CAD systems for colorectal cancer from WSI are still not ready for clinical acceptance. *Scientific Reports* 11.1 (2021): 14358.
11. Wei, Jerry, et al. A petri dish for histopathology image analysis. *Artificial Intelligence in Medicine: 19th International Conference on Artificial Intelligence in Medicine, AIME 2021, Virtual Event, June 15–18, 2021, Proceedings*. Springer International Publishing, 2021.
12. Mota, Tiago, et al. "MMIST-ccRCC: A Real World Medical Dataset for the Development of Multi-Modal Systems." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.
13. Oquab, Maxime, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).
14. He, Kaiming, et al. "Masked autoencoders are scalable vision learners." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.