**Unlearning Images from a Pre-trained Model : Erase the influence of requested samples without hurting accuracy**

Hema Thirumala Prasad Kalangi

Hon Lam Chung

Rohini Sidharth Kulkarni

Shreya Singh

Department of Applied Data Science, San Jose State University

DATA255: Deep Learning Technologies

Shim, Simon S.Y.

**Abstract**

Deep learning models have been widely adopted, but this has raised privacy concerns, that solves the problem of extracting data from a trained model without sacrificing much accuracy. This is a difficult problem because algorithms must be able to forget specific information while retaining the same level of performance. Unlearning became more significant because it aims to filter out sensitive information from the trained model. It is mostly used in privacy protection, bias mitigation and security improvement. For example, by unlearning a set of human faces from a facial recognition model, those people’s privacy who are “forgotten” by the algorithm are now protected. Unlearning can also be used to reduce the potential bias in AI models . This project aims to implement machine unlearning on a pre-trained age predictor. To un-educate the model, an unlearning algorithm is required to instill the impact of the forget set and minimize the original image’s information. The main motivation behind this project is privacy protection. In this case, the unlearned model will be used to protect the privacy of the individuals. Two approaches are used in terms of designing the unlearning workflow. The first approach is called Gradient-based unlearning, which works by iteratively updating a model's weights in such a way that minimizes the influence of the forgotten images. The second approach is to use adversarial unlearning techniques. It works by training an adversarial model to generate images that are similar to the forgotten images. The adversarial model is then used to train the original model to be less sensitive to the forgotten images.We will use the SISA with CNN and Siamese Networks to implement neurons unlearn unlearning algorithm approach for different neural models. Dataset will be provided in this kaggle competition by the organizers.

**Dataset links and description**

The dataset consists of multiple components. First and foremost, it has a set of images that the pre-trained model needs to forget. It comes with a metadata file that contains 4 features: person\_id, age\_group, age, and image\_id. The combination of person\_id and image\_id is a unique key to differentiate each image. The dataset also has a JSON file of age’s class weight from the pre-trained model. Last but not least, a model checkpoint in PyTorch for the pre-trained model.

“<https://www.kaggle.com/competitions/neurips-2023-machine-unlearning/data>”

**Technologies**

Google colaboratory will be the main IDE for this project to ensure all the Python libraries used are the same version. Depending on the model complexity, a different runtime type might be required, such as T4 GPU, so that the untraining part does not take forever to complete. For data wrangling and transformation within the machine unlearning algorithm, numpy and Scikit Learn will be major contributors as both libraries are very good at transforming and evaluating data.

**Reference links:**

Nguyen, T. T., Huynh, T. T., Nguyen, P. L., Liew, A. W. C., Yin, H., & Nguyen, Q. V. H. (2022). A survey of machine unlearning. arXiv preprint arXiv:2209.022

[arXiv:2209.02299v5](https://arxiv.org/abs/2209.02299v5)

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., ... & Papernot, N. (2021, May). Machine unlearning. In 2021 IEEE Symposium on Security and Privacy (SP) (pp. 141-159). IEEE.

[arXiv:1912.03817v3](https://arxiv.org/abs/1912.03817v3)

Xu, H., Zhu, T., Zhang, L., Zhou, W., & Yu, P. S. (2023). Machine unlearning: A survey. ACM Computing Surveys, 56(1), 1-36.

<https://dl.acm.org/doi/pdf/10.1145/3603620?casa_token=he3Nqx0uWR8AAAAA:NnB2VCohOp0Cf4OdStPGHEcPmQpOTe6D4geqnpBg-BNwHVVx0x4-Mh6tdhLCkbXvp_n5xqY-xML->