MINT-Demo: Membership Inference Test Demonstrator

Daniel DeAlcala, Aythami Morales, Julian Fierrez, Gonzalo Mancera, Ruben Tolosana, Ruben Vera-Rodriguez

Biometrics and Data Pattern Analytics Lab, Universidad Autonoma de Madrid, Spain

Abstract

We present the Membership Inference Test Demonstrator, to emphasize the need for more transparent machine learning training processes. MINT is a technique for experimentally determining whether certain data has been used during the training of machine learning models. We conduct experiments with popular face recognition models and 5 public databases containing over 22M images. Promising results, up to 89% accuracy are achieved, suggesting that it is possible to recognize if an AI model has been trained with specific data. Finally, we present a MINT platform as demonstrator of this technology aimed to promote transparency in AI training[1].

Introduction

The unauthorized use of personal or copyrighted material to train AI models may infringe upon the rights of individuals. Moreover, the generated output of AI models trained on this data may blur the line between original and derived works, raising concerns of plagiarism and copyright infringement.

On June 2023, the European Parliament adopted its negotiating position on the Artificial Intelligence Act (European Commission 2023) requiring AI providers to ensure robust protection of fundamental citizen rights. The regulation mandates the registration of AI models in an EU database and grants national authorities the power to request access to trained models and their training data. This regulation enforces transparency and calls for new auditing tools to ensure secure AI deployment in Europe.

These considerations lead us to the main objective of this work, which is to propose a platform to detect the data used to train AI models. Currently, developers can hide behind the weights of their network to bypass regulations and conceal the use of training data from users.This approach seeks to unveil AI training processes, ensuring alignment with legislation and citizen rights.

The main contributions can be summarized as follows:

- Introduced MINT, a method to detect data usage during AI training (Fig. 1), aiding compliance with AI legislation and protecting citizen rights.
- Conducted experiments on 5 public datasets with 22M+ images, achieving up to 89% accuracy in Membership Inference Tests, highlighting its challenges and benefits.
- Developed an interactive web platform with real models to promote AI transparency.

[1]https://ai-mintest.org/

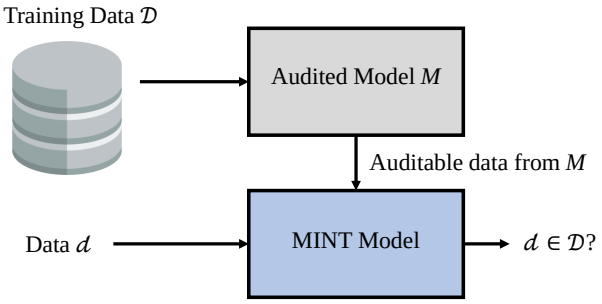

Figure 1: Block diagram of the Membership Inference Test.

Membership Inference Test

Let us consider a Training Dataset $\mathcal{D}$, an External Dataset $\mathcal{E}$ and a collection of samples $d \in \mathcal{D} \cup \mathcal{E}$. We assume a learned model M that is trained for a specific task (text generation, face recognition, etc.) using the dataset $\mathcal{D}$. For any input data record d, the model M generates an outcome y based on d and a set of parameters $\mathbf{w}$ ($y = M(d|\mathbf{w})$) and intermediate outcomes or Auxiliary Auditable Data AAD (e.g., activation maps of specific layers in a Neural Network) based on d and a subset of parameters $\mathbf{w}'$ ($AAD = N(d|\mathbf{w}')$).

The Membership Inference Test (MINT) aims to determine if a data d was used to train the model M, i.e., if d belongs to the training dataset $\mathcal{D}$ or External Data $\mathcal{E}$ ($\mathcal{E} \notin \mathcal{D}$). To this end, an authorized authority employs the final and intermediate outcomes to train an auditing model ($T(\cdot|\theta)$). These terms can be seen within the entire workflow in Fig 2.

Membership Inference Test: Experiments

We present the experiments with a popular face recognition model from the InsigthFace project (InsightFace Team 2023). However, the same experiments have been conducted with other face reconition models and can be tested on the website (https://ai-mintest.org/). The face recognition model used (M in Fig. 1) is a ResNet-100 network (Han, Kim, and Kim 2017), trained on the Glint360k database (An et al.

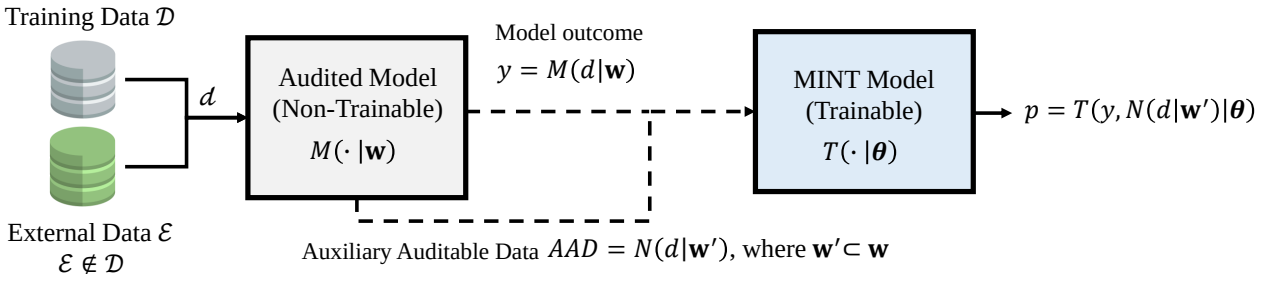

Figure 2: The MINT Model (T) predicts whether specific data (d) was used to train an Audited AI Model (M), using Auxiliary Auditable Data (e.g., activation maps) and/or the model outcome from M.

2021) with CosFace loss function (Wang et al. 2018). This database comprise 17M images ($\mathcal{D}$ in Fig. 1).

We propose two different MINT model architectures:

1. **Vanilla MINT Model**: An MLP consisting of three fully connected layers—input-size neurons (varying by Auxiliary Auditable Data), 64 neurons, and 1 neuron. A dropout layer (0.3 rate) and an L1 regularizer (0.1) are applied between layers.
2. **CNN MINT Model**: A CNN with two convolutional layers (64 and 128 filters) followed by two fully connected layers sized to the convolution output.

We included the IJB-C (Maze et al. 2018), FDDB (Jain and Learned-Miller 2010), GANDiffFace (Melzi et al. 2023), and Adience (Eidinger, Enbar, and Hassner 2014) databases as external Data ($\mathcal{E}$) to train and test the MINT model T. For the Auxiliary Auditable Data, we used activations from various layers in M. In the Vanilla MINT Model, we extract the maximum value from each activation map at different depths, forming a vector whose size depends on the number of filters in the selected layer. The CNN MINT Model, however, analyzes activations directly, using the full activation maps without vectorizing them.

Table 1 presents the classification accuracy for the Vanilla MINT model. The columns represent the number of samples used to train the MINT model (T), and the rows show the depth of the selected activation maps (Auxiliary Auditable Data). We focus on the final convolutional layer of each of the 4 major ResNet-100 blocks (First to Fourth layers). The table also includes the "output layer" (model's output embedding) and "all conv layers" (concatenated Conv Layers). The classification accuracy varies depending on the available Auxiliary Auditable Data and amount of data, with "all layers" yielding the best performance (up to 84%). The best individual results come from layers closest to the input and output, while intermediate layers show poorer outcomes.

Table 2, presents the results for the CNN MINT Model. Notably, there are no results for the Model Outcome, as CNN architectures cannot be applied directly to the output vector. Similarly, the row for concatenating convolutional layers is missing due to the varying resolutions of activation maps, making concatenation impractical—unlike the Vanilla MINT Model where vectorization made it feasible. In this architecture, the best performance is achieved with the layer closest to the input, with accuracy decreasing towards the output. The CNN MINT Model achieves 89% accuracy, outperforming the Vanilla model's 84%.

Auditable Data	1K samples	50K samples	100K samples
Conv Layer #1	0.62	0.80	0.80
Conv Layer #2	0.56	0.67	0.68
Conv Layer #3	0.56	0.58	0.59
Conv Layer #4	0.73	0.76	0.76
Model Outcome	0.67	0.78	0.78
All Conv Layers	**0.76**	**0.82**	**0.84**

Table 1: Classification accuracy using the Vanilla MINT Model. The MINT model was trained with a variable number of samples ranging from $100K$ to $1k$.

Auditable Data	1K samples	50K samples	100K samples
Conv Layer #1	**0.88**	**0.89**	**0.89**
Conv Layer #2	0.85	0.86	0.86
Conv Layer #3	0.68	0.71	0.75
Conv Layer #4	0.68	0.70	0.74

Table 2: Classification accuracy using the CNN MINT Model. The MINT model was trained with a variable number of samples ranging from $100K$ to $1k$.

Membership Inference Test: Demonstrator

In this work we introduce the MINT web platform, https://ai-mintest.org/. On this platform, citizens can upload images and receive reports on the likelihood that these images were used to train an AI model. This demonstrator includes a limited initial set of popular models. Designed to promote transparency in AI, the platform will expand to include more models and across various data types (e.g., text, audio, image). This platform opens up new opportunities for research and encourages the development of tools, standards, and protocols to comply with the new regulations.

Acknowledgement

This work has been supported by projects BBfor-TAI (PID2021-127641OB-I00 MICINN/FEDER), Cátedra ENIA UAM-VERIDAS (NextGenerationEU PRTR TSI-100927-2023-2), and Comunidad de Madrid (ELLIS Unit Madrid). The work of D. DeAlcala is supported by a FPU Fellowship (FPU21/05785) from the Spanish MIU.

References

An, X.; Zhu, X.; Gao, Y.; Xiao, Y.; Zhao, Y.; Feng, Z.; Wu, L.; Qin, B.; Zhang, M.; Zhang, D.; et al. 2021. Partial fc: Training 10 million identities on a single machine. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1445–1449.

Eidinger, E.; Enbar, R.; and Hassner, T. 2014. Age and Gender Estimation of Unfiltered Faces. *IEEE Transactions on Information Forensics and Security*, 9(12): 2170–2179.

European Commission. 2023. Artificial Intelligence Act. *EU Legislation in Progress*.

Han, D.; Kim, J.; and Kim, J. 2017. Deep pyramidal residual networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5927–5935.

InsightFace Team. 2023. InsightFace Models. *https://insightface.ai/*.

Jain, V.; and Learned-Miller, E. 2010. Fddb: A benchmark for face detection in unconstrained settings. Technical report, UMass Amherst technical report.

Maze, B.; Adams, J.; Duncan, J. A.; Kalka, N.; Miller, T.; Otto, C.; Jain, A. K.; Niggel, W. T.; Anderson, J.; Cheney, J.; and Grother, P. 2018. IARPA Janus Benchmark - C: Face Dataset and Protocol. In *Proceedings of the International Conference on Biometrics*, 158–165.

Melzi, P.; Rathgeb, C.; Tolosana, R.; Vera-Rodriguez, R.; Lawatsch, D.; Domin, F.; and Schaubert, M. 2023. GANDiffFace: Controllable Generation of Synthetic Datasets for Face Recognition with Realistic Variations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*.

Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; and Liu, W. 2018. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5265–5274.