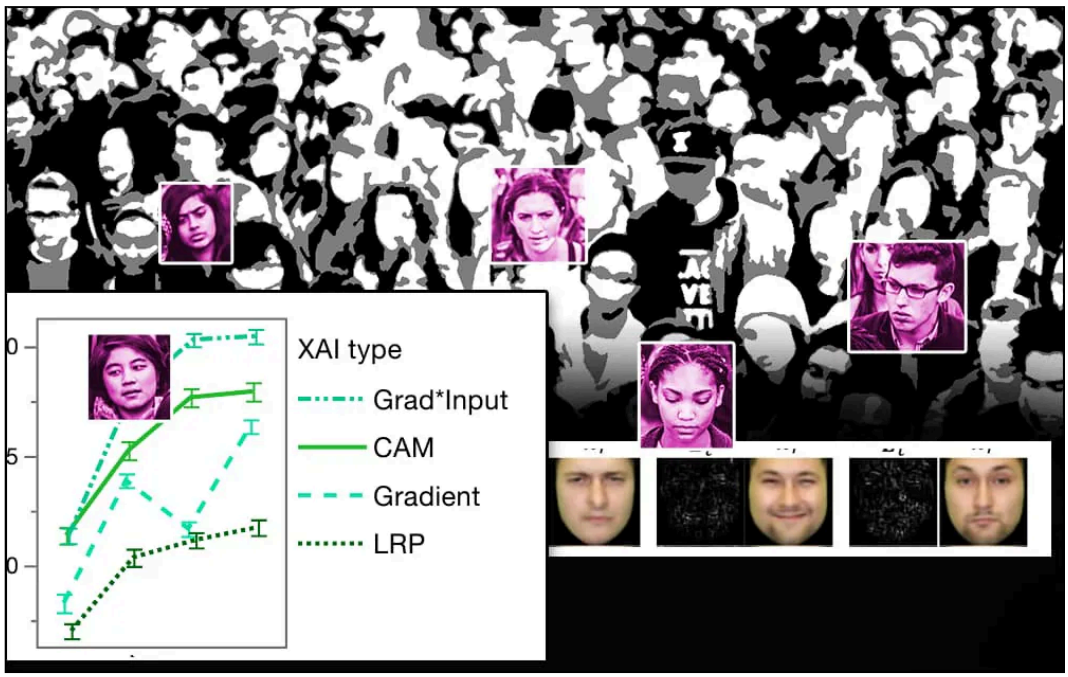
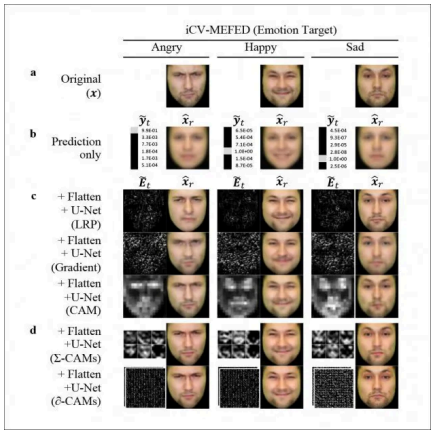
# Explainable AI May Surrender Confidential Data More Easily

Originally published at https://www.unite.ai/explainable-ai-may-surrender-confidential-data-more-easily/




Researchers from the National University of Singapore have concluded that the more explainable AI becomes, the easier it will become to circumvent vital privacy features in machine learning systems. They also found that even when a model is not explainable, it's possible to use explanations of similar models to 'decode' sensitive data in the non-explainable model.

The research, titled *Exploiting Explanations for Model Inversion Attacks*, highlights the risks of using the 'accidental' opacity of the way neural networks function as if this was a by-design security feature – not least because a wave of new global initiatives, including the European Union's draft AI regulations, are characterizing explainable AI (XAI) as a prerequisite for the eventual normalization of machine learning in society.




*In the research, an actual identity is successfully reconstructed from supposedly anonymous data relating to facial expressions, through the exploitation of multiple explanations of the machine learning system.* Source: https://arxiv.org/pdf/2108.10800.pdf

The researchers comment:

> *'Explainable artificial intelligence (XAI) provides more information to help users to understand model decisions, yet this additional knowledge exposes additional risks for privacy attacks. Hence, providing explanation harms privacy.'*

## Re-Identification of Private Data

Participants in machine learning datasets may have consented to be included on the assumption of anonymity; in the case of Personal Identifiable Information (PII) that ends up in AI systems via ad hoc data-gathering (for instance, through social networks), participation may be technically legal, but strains the notion of 'consent'.

Several methods have emerged in recent years that have proved capable of de-anonymizing PII from apparently opaque machine learning data flows. Model extraction uses API access (i.e. 'black box' access, with no special availability of the source code or data) to extract PII even from high-scale MLaaS providers, including Amazon Web Services, while Membership Inference Attacks (MIAs), operating under similar constraints, can potentially obtain confidential medical information; additionally Attribution Inference Attacks (AIAs) can recover sensitive data from API output.

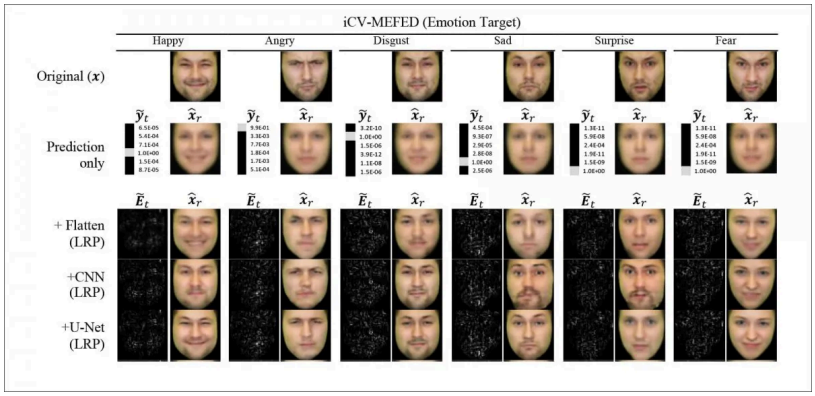
## Revealing Faces

For the new paper, the researchers have concentrated on a model inversion attack designed to obtain an identity from a subset of facial emotion data that should not be capable of revealing this information.

The objective of the system was to associate images found in the wild (either posted casually on the internet or in a potential data breach) with their inclusion in the datasets that underpin a machine learning algorithm.

The researchers trained an inversion attack model capable of reconstructing the contributing image from the anonymized API output, without special access to the original architecture. Previous work in this field has concentrated on systems where identification (protecting or revealing) was the objective of both the target system and the attacking system; in this case, the framework has been designed to exploit the output of one domain and apply it to a different domain.

A transposed convolutional neural network (CNN) was employed to predict an 'original' source face based on the target prediction vector (saliency map) for an emotion recognition system, using a U-Net architecture to improve facial reconstruction performance.




*The re-identification system is powered and informed by explainable AI (XAI), where knowledge of neuron activation, among many contributing public XAI facets, i internal machinations of the architecture only from its output, enabling re-identification of contributing dataset images.*

## Testing

In testing the system, the researchers applied it against three datasets: iCV-MEFED face expressions; CelebA; and MNIST handwritten digits. To accommodate the model size being used by the researchers, the three datasets were resized respectively to 128×128, 265×256 and 32×32 pixels. 50% of each set was used as training data, and the other half used as an attack dataset to train the antagonist models.

Each dataset had different target models, and each attack network was scaled to the limitations of the explanations underpinning the process, rather than using deeper neural models whose complexity would exceed the generalization of the explanations.

The XAI explanation types used to power the attempts included Gradient Explanation, Gradient Input, Grad-CAM and Layer-Wise Relevance Propagation (LRP). The researchers also evaluated multiple explanations across the experiments.

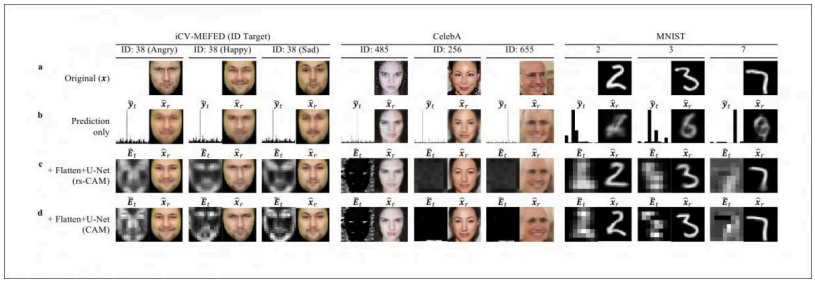



*Image reconstruction facilitated by an XAI-cognizant inversion attack across the three datasets, featuring identical target and attack tasks.*

The metrics for the test were pixelwise similarity evaluated by Mean Squared Error (MSE); Image Similarity (SSIM), a perceptually-based similarity index; attack accuracy, determined by whether a classifier can successfully re-label a reconstructed image; and attack embedding similarity, which compares the feature embeddings of known source data against reconstructed data.

Re-identification was achieved, with varying levels according to the task and the datasets, across all the sets. Further, the researchers found that by concocting a surrogate target model (which they naturally had complete control over), it was still possible to achieve re-identification of data from external, 'closed' models, based on known XAI principles.

The researchers found that the most accurate results were obtained by activation-based (saliency map) explanations, which leaked more PII than sensitivity-based (gradient) approaches.

In future work, the team intends to incorporate different types of XAI explanation into novel attacks, such as feature visualizations and concept activation vectors.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

**Discover More**