Exploiting Explanations for Model Inversion Attacks

Xuejun Zhao Wencan Zhang Xiaokui Xiao Brian Lim[*]
National University of Singapore, Singapore
{xuejunzhao, wencanz}@u.nus.edu, xkxiao@nus.edu.sg, brianlim@comp.nus.edu.sg

Abstract

The successful deployment of artificial intelligence (AI) in many domains from healthcare to hiring requires their responsible use, particularly in model explanations and privacy. 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. We study this risk for image-based model inversion attacks and identified several attack architectures with increasing performance to reconstruct private image data from model explanations. We have developed several multi-modal transposed CNN architectures that achieve significantly higher inversion performance than using the target model prediction only. These XAI-aware inversion models were designed to exploit the spatial knowledge in image explanations. To understand which explanations have higher privacy risk, we analyzed how various explanation types and factors influence inversion performance. In spite of some models not providing explanations, we further demonstrate increased inversion performance even for non-explainable target models by exploiting explanations of surrogate models through attention transfer. This method first inverts an explanation from the target prediction, then reconstructs the target image. These threats highlight the urgent and significant privacy risks of explanations and calls attention for new privacy preservation techniques that balance the dual-requirement for AI explainability and privacy.

1. Introduction

The recent success of artificial intelligence (AI) is driving its application in many domains from healthcare to hiring. With the increasing regulatory requirements for responsible AI, these applications will need to support both explainability to justify important decisions and privacy to protect personal information [36]. Indeed, many AI applications have this dual-requirement, such as driver drowsiness detection from faces [13, 35] that can be used to limit the jobs of tired drivers, and classroom facial engagement prediction [6, 49] for student grading or teacher performance evaluation. Users need explanations to dispute or rectify unfavorable model predictions, and need privacy to preserve anonymity or reputation (e.g., of embarrassing appearances). Troublingly, using state-of-the-art model inversion attacks [11, 50, 54], attackers can reconstruct sensitive information (e.g., faces) merely from model predictions. We hypothesize that since explanations provide more information, they can be used by attackers for more effec-

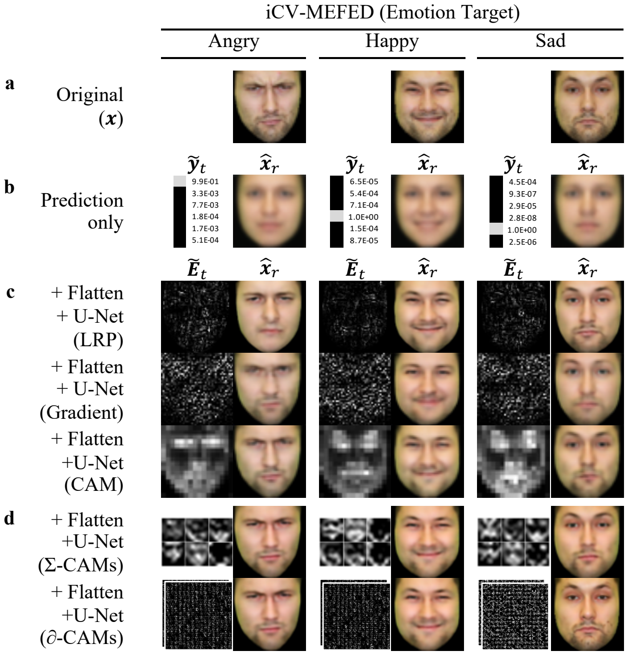

Figure 1. Demonstration of image reconstruction from XAI-aware inversion attack with emotion prediction as target task, and face reconstruction as attack task. Three emotions from a single identity shown from the iCV-MEFED dataset [28]. Reconstructions shown with corresponding inputs (target prediction $\tilde{y}_t$ and explanations $\tilde{E}_t$ as LRP [3],Gradients [43] or Grad-CAM [39] saliency maps). Towards original images (a), reconstructions from Prediction only (b) are poor and similar across different faces, and is significantly improved when exploiting single (c) and multiple (d) explanations. Demonstration reconstructions for other explanations and baseline inversion models are shown in supplementary materials.

[*]Corresponding author.

tive reconstructions, and further harm privacy — providing explanations harms privacy.

Explainable artificial intelligence (XAI) provides information to help users to debug models [23], improve decision making [47], and improve and moderate trust in automation [10]. Specific to image-based deep learning, XAI techniques include saliency maps [7, 34, 43, 39, 56], feature visualization [4, 16, 32], and activations of neurons [18] and concept vectors [19, 20]. These explanations provide users with deeper insights into model reasoning and about the data, but they can contain sensitive knowledge that can be exploited for privacy attacks, especially with recent machine learning (ML)-based attacks [41, 46, 11]. We focus on model inversion attacks that can reconstruct data from model predictions [11, 15, 50, 54]. These methods typically require some privileged white-box access (to model parameters) or background knowledge (such as blurred images) for effective attack, but this is unrealistic or uncommon in deployed settings. Instead, because of the requirement for explainability, model explanations will be more readily available and pose a more ubiquitous threat.

Recent work has begun to study the privacy threats of model explanations, but their attack goals differ and their use of explanations remains under-studied. Instead of attacking training data [41] or model confidentiality [2, 30], we focus on attacking the privacy of test instances. This will compromise the trust of active users of the model. We propose attack methods based on the structure and generation technique of explanations, focusing on saliency maps that are highly popular for image-based AI. Furthermore, we propose a method of leveraging on XAI techniques with attention transfer to perform inversion attacks on models that do not share explanations. We found that this performance is close to inverting an explainable target model. Hence, even non-explainable target models are at increased risk of XAI-aware inversion attacks.

Our contributions are: **1)** We determine the privacy threat of model explanations for model inversion attacks. We achieved significantly higher inversion performance by proposing several architectures for XAI-aware model inversion, through careful adapting of multi-modal, spatially-aware transposed CNNs. This highlights that the present privacy risk of providing explanations is significant (see Figure 1). **2)** We propose an XAI-aware attack on non-explainable target models that can achieve improved inversion performance using an explanation inversion model with surrogate explanation attention transfer. This does not need additional data leak or sharing from the target model, and demonstrates a significantly increased risk due to explanations in surrogate models instead of the target model. **3)** We identify the privacy risk for different explanation types (Gradients [43], Grad-CAM [39], and layer-wise relevance propagation (LRP) [3]) and analyze how various factors influence attack performance.

In this work, we provide the first study into the privacy risk of explanations for inversion attack, and defer their defense to future work. This highlights the urgency to develop privacy defense techniques and models that are co-optimized for the dual-requirements of explainability and privacy to achieve the objectives of Responsible AI.

2. Related Works

Our work relates to two research areas — explainable AI and machine learning-based privacy — which we overview separately, and discuss nascent work intersecting the two.

Explainable AI for image-based CNN models. While there are many XAI techniques (see review [5, 14]), we focus on explanations of image-based convolutional networks (CNN) (see review [53]) and specifically saliency map explanations that highlight important pixels for each prediction [7, 34, 43, 39, 56]. Gradients [43] of the model prediction with respect to input features, i.e., pixels in an image, describe the model sensitivity. Though easy to calculate, this suffers from several issues, which are addressed in extensions, such as robustness (SmoothGrad [44]), misleading saturated gradients (DeepLIFT [42]), and implementation invariance (Integrated Gradients [45]). For this work, we study the exploitation of the general gradient technique [43], and expect that our findings will generalize to the extensions. Strictly, these techniques produce sensitivity maps and describe how much the prediction may change if the pixels are altered. Alternatively, activation map explanations show how influential each pixel for the prediction. One simple approach is Gradient $\odot$ Input [42] which is the element-wise product of gradient and image to approximate the model activation based on the image, but this highlights fine-grained details that may be hard to read.

Class activation maps [7, 34, 39, 56] provide class-specific coarse-grained explanations based on activation maps of convolution filters. While CAM [56] required model retraining, Grad-CAM [39] could be read from any CNN without retraining. Extensions improved explanations for robustness (Ablation-CAM [34]) and to handle multiple objects (Grad-CAM++ [7]). For this work, we evaluated with the original Grad-CAM [39] approach, and expect that our findings will generalize to its extensions. While the aforementioned techniques are model-specific and require white box access to model parameters, model-agnostic approaches can apply generally to any deep neural network, such as perturbation-based sensitivity analysis [52], and layer-wise relevance propagation (LRP) [3]. However, as proxy explanations, these techniques are less faithful to the model behavior, and we defer their study.

Other than helping user understanding, saliency map explanations have also been used to improve model training through transfer learning from a better model with student-teacher networks [21]. Attention can also be indirectly transferred by maximizing the loss between predicting on an image and its variant ablated with its saliency mask (GAIN [25]). Similarly, to attack non-explainable target models, we exploit attention transfer to minimize the intermediate explanation in the attack model and the explanation of the surrogate target model.

Machine learning privacy and model inversion attacks. Recent research on machine learning privacy has identified sophisticated attacks, such as model extraction attacks [46, 38] to reconstruct parameters of proprietary models, membership inference attacks [41] to identify if users were part of a training dataset (e.g., of cancer patients [17]), attribution inference attacks [12] to impute omitted or obfuscated data to recover the sensitive information, and model inversion attacks [11, 50, 54] to infer the original data from the target prediction (e.g., reconstructing a face from an emotion prediction). In this work, we focus on the latter attack, which is relevant to image-based machine learning, where private input images can be recovered, thus de-anonymizing users or revealing sensitive details. While the original model inversion attack has limited performance [11], inversion performance is notably improved by leveraging deep architectures [15, 50] (specifically, transposed CNNs [8]). Further improvements exploit auxiliary knowledge, such as white-box model access [54], feature embeddings [15], blurred images [54], or the joint probability distribution of features and labels [51], but such information is not readily available in practice. However, the need for explanations will increase their ubiquity and this will poses increased privacy risk.

Privacy risk of model explanations. Membership inference attacks can be improved by concatenating gradient explanations with model predictions [40]. Model extraction attacks can be improved by regularizing the reconstructed model with gradient explanations [30], and training a surrogate model from counterfactual explanation examples [2]. However, exploiting explanations for model inversion attacks remains unexplored; this conceals the privacy risk on user data at prediction time, i.e., of active users. In this work, we investigate how to exploit different saliency map explanations types to attack explainable and non-explainable models for inversion attacks.

3. XAI-Aware Model Inversion Attack

We describe the general approach for model inversion and describe how to exploit explanations for more aggressive attacks through various architectures (see Figure 2).

3.1. Threat Model

Consider a target model (M_t) that has been trained and is deployed for use through an application programming interface (API). It takes private data $\boldsymbol{x} \in X_p$ (e.g., face image) provided by a user to produce a *target prediction* $\tilde{\boldsymbol{y}}_t$ (e.g., emotion) as a confidence probability vector. It is trained with cross-entropy loss $L_t(\boldsymbol{y}_t, \tilde{\boldsymbol{y}})$, where $\boldsymbol{y}_t$ is the ground truth label for the target task. To improve user trust and verification, the target model also provides a *target explanation* $\tilde{\boldsymbol{E}}_t$ for each prediction (e.g., saliency map [39]). Consider a model inversion attack at run time, where the attacker obtains access to each tuple of the target prediction vector and explanation tensor (e.g., due to breached storage, intercepted in transit, shared via social media). We assume that the attacker only has access to the breached data, an independent dataset $\boldsymbol{x} \in X_a$, and the target model via its API (i.e., black-box access, unlike white-box access in [11, 54]). Furthermore, unlike [50, 54], we do not require privileged background knowledge (e.g., blurred images). The attack goal is to train an inversion attack model M_i^a to reconstruct the original image $\boldsymbol{x}$ from the model outputs $(\tilde{\boldsymbol{y}}_t, \tilde{\boldsymbol{E}}_t)$ such that sensitive information can be predicted from the reconstructed image $\hat{\boldsymbol{x}}_r$. Using facial emotion recognition as the

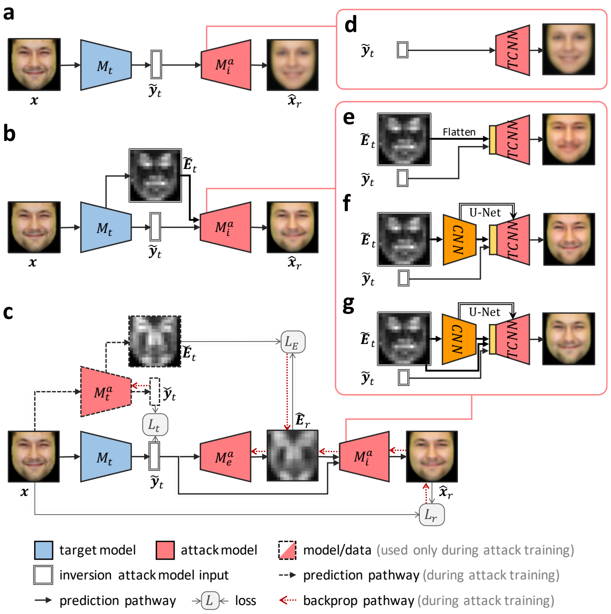
a
b
c
d
e
f
g
$\boldsymbol{x}$
M_t
$\tilde{\boldsymbol{y}}_t$
M_i^a
$\hat{\boldsymbol{x}}_r$
$\tilde{\boldsymbol{E}}_t$
M_t^a
M_e^a
L_E
L_t
L_r
$\hat{\boldsymbol{E}}_r$
TCNN
Flatten
U-Net
CNN
target model
attack model
model/data (used only during attack training)
inversion attack model input
prediction pathway (during attack training)
prediction pathway
loss
backprop pathway (during attack training)

Figure 2. Architectures of inversion attack models. a) Baseline threat model with target CNN model M_t to predict emotion $\tilde{\boldsymbol{y}}_t$ from face $\boldsymbol{x}$, and inversion attack model to reconstruct face $\hat{\boldsymbol{x}}_r$ from emotion. Emotion prediction confidences are input to a transposed CNN (TCNN) for inversion attack (d). b) Threat model with explainable target model that also provides instance explanation $\tilde{\boldsymbol{E}}_t$ of the target prediction, and XAI-aware multi-modal inversion attack model that inputs $\tilde{\boldsymbol{E}}_t$ via different input architectures: e) Flattened $\tilde{\boldsymbol{E}}_t$ concatenated with $\tilde{\boldsymbol{y}}_t$, f) U-Net for dimensionality reduction and spatial knowledge, g) combined Flatten and U-Net. Additional input architectures shown in supplementary materials. c) Threat model with non-explainable target model, and inversion attack model that predicts a reconstructed surrogate explanation $\hat{\boldsymbol{E}}_r$ from target prediction $\tilde{\boldsymbol{y}}_t$ and uses $\hat{\boldsymbol{E}}_r$ for multi-modal image inversion (e-g).

running example for the target model, the attack goal is to reconstruct the face image and re-identify the user. This poses problems of consent and identity theft. In our case, the target and attack tasks are not identical; this differs from prior works that have identity prediction for both target and attack tasks [11, 54, 50]. Figure 2a illustrates the baseline model inversion attack using only target model predictions. We will next describe more serious attacks by exploiting target model explanations.

3.2. Model Inversion with Target Explanations

To invert the target model M_t, [50], we trained the inversion attack model M_i^a as a Transposed CNN (TCNN) [9] to predict a 2D image $\hat{\boldsymbol{x}}_r$ from the 1D target prediction vector $\tilde{\boldsymbol{y}}_t$ as input to the attack model (see Figure 2a). M_i^a is trained with MSE loss for the image reconstruction $L_r(\boldsymbol{x}, \hat{\boldsymbol{x}}_r)$. We illustrate the convolutional layers for upsampling in the TCNN in the supplementary materials. We consider saliency map explanations (e.g., [43, 39]) and extend the TCNN model to be multi-modal to include the 2D target explanation as a second input (Figure 2b). A simple approach is to flatten the explanation into a 1D vector, concatenate with the prediction vector and input to the TCNN (Figure 2e). However, this neglects the spatial information of the salient pixels. A simple way to leverage the 2D information is with a CNN architecture that uses convolutional kernels to interpret 2D patterns into a 1D feature embedding vector which is put into the TCNN (see supplementary materials). This CNN-TCNN architecture is similar to the generator proposed in [54]. Inspired by CNN encoder-decoder networks [33] and image super-resolution techniques [55], we propose to use a U-Net architecture [37] in the inversion model to improve its reconstruction performance (Figure 2f). We hypothesize that this will have more accurate reconstruction because of the retention in spatial information that is transmitted through the bypass connectors from the convolutional layers. We further propose a hybrid model by combining the Flatten and U-Net input architectures as the Flatten+U-Net inversion model (Figure 2g); the 1D flattened explanation and prediction vectors are concatenated with the latent layer of the U-Net architecture. The proposed architectures demonstrate various methods to obtain sensitive information for privacy attack; we heretofore call them XAI Input Methods. The training loss function regards image reconstruction task:

$$L_r = \sum_x (M_i^a(M_t(\boldsymbol{x})) - \boldsymbol{x})^2 \tag{1}$$

where $\boldsymbol{x}$ is the private input image, $M_t(\boldsymbol{x}) = \tilde{\boldsymbol{y}}_t$ is the target model prediction, and $M_i^a(M_t(\boldsymbol{x})) = \hat{\boldsymbol{x}}_r$ is the reconstructed image from the inversion attack.

Later, we describe our experiments on how some architectures capture more knowledge from the explanations and interpret why. The proposed architectures can be used for any 2D explanations of the CNN target model, such as gradients [43], class activation maps (CAM) [56], LRP [3], and 2D auxiliary data, such as blurred variants of the input images x_b. We further conducted experiments on different Explanation Types, described later.

3.3. Model Inversion with Multiple Explanations

While many explanations explain *why* a model predicted a class $c \in C$, it is also important to explain *why not* an alternative class $c' \neq c$, i.e., to provide contrastive explanations [26, 29]. To support this, some explanation techniques, such as Grad-CAM [39], can provide class-specific explanations depending on user query. However, this further raises the privacy risk, since more information is provided by the additional explanations. We exploit these Alternative CAMs (Σ-CAM) by concatenating explanations for $|C|$ classes into a 3D tensor and training the inversion models on this instead of the 2D matrix of a single explanation. Supplementary Figure 1f for detailed architecture.

To gain a deeper insight into a model decision, users may be interested in understanding which specific neurons of a CNN were activated and how. In particular, although Grad-CAM presents one saliency map per class c, this is composed as a weighted sum of activation maps from the last convolutional layer in the CNN, i.e., the saliency map explanation is $E^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right)$, where A^k is the kth activation map ($k \in K$), and $\alpha_k^c = \frac{1}{HW} \sum_{i,j}^{H,W} \frac{\partial y^c}{\partial A_{ij}^k}$ is the gradient-based importance weight. Hence, the Grad-CAM explanation is composed of $|K|$ partial, constituent CAMs (∂-CAM). Combined with explanation techniques to understand the role of each neuron (e.g., [4, 32]), these ∂-CAMs can provide rich insights for developers. However, since $|K| \gg 1$ is large, this provides a multitude of explanations that further leak privacy. Similar to Alternative CAMs, we exploit these Constituent CAMs ∂-CAM by concatenating them into a 3D tensor as input to the inversion model.

3.4. Model Inversion with Surrogate Explanations

While the XAI Input methods can increase privacy risk, we further hypothesize that XAI techniques can be exploited for inversion attacks even for non-explainable target models (i.e., no target explanation). We propose an architecture that predicts the target explanation and exploits that explanation to invert the original target data. Figure 2c illustrates the architecture for this attack. First, we train an explainable surrogate target model M_t^a on the attacker's dataset to generate surrogate explanation $\check{\boldsymbol{E}}_t$. Then, as $X_a \to X$, then $M_t^a \to M_t$ and $\check{E} \to \tilde{E}$.

However, $\check{\boldsymbol{E}}_t$ is only available during training not when predicting. Hence, we train an explanation inversion model M_e^a to reconstruct $\check{\boldsymbol{E}}_t$ as $\hat{\boldsymbol{E}}_r$ from the target prediction $\tilde{\boldsymbol{y}}_t$

by minimizing the surrogate explanation loss function:

$$L_E = \sum_x (M_e^a(M_t(\boldsymbol{x})) - E(M_t^a(\boldsymbol{x})))^2 \tag{2}$$

where $E(M)$ is the explanation of the model M, $M_t^a(\boldsymbol{x}) = \check{\boldsymbol{y}}_t$ is the surrogate target prediction, $E(M_t^a(\boldsymbol{x})) = \check{\boldsymbol{E}}_t$ is the surrogate explanation, and $M_e^a(M_t(\boldsymbol{x})) = \hat{\boldsymbol{E}}_r$ is the reconstructed surrogate explanation. This reconstructed explanation is available at prediction time. Finally, we input $\hat{\boldsymbol{E}}_r$ into the image inversion model M_i^a to complete the model inversion attack. Since $\hat{\boldsymbol{E}}_r$ is the same format as $\tilde{\boldsymbol{E}}_t$, we can apply any of the aforementioned XAI Input Methods. Overall, the proposed architecture can also be described as model inversion attack with attention transfer, where explanations are transferred from surrogate target model M_t^a into the intermediate layer between M_e^a and M_i^a. This guides and constrains the inversion model to learn activations that M_t^a finds relevant for $\check{y}$ as a proxy for $\hat{y}$. The training loss involves two tasks: image reconstruction (Eq. 1) and explanation reconstruction (Eq. 2).

4. Experiments

We conducted experiments to perform an ablation analysis of our model inversion architectures and to compare with existing work. We evaluated on multiple datasets for different use cases and evaluated image reconstruction quality and accuracy to classify sensitive information from reconstructed images. Figure 3 demonstrates increased inversion risk due to explanations and XAI-aware attention transfer.

4.1. Experiment Setup

Use Cases and Datasets. We evaluated on three datasets representing three use cases with the simultaneous need for model explanations and privacy. 1) iCV-MEFED face expressions data [28] with 6 emotions and 115 identities (10 omitted due to confidentiality) over 24,120 instances. We evaluated a first use case where an explainable target model predicts emotion from faces (e.g., for classroom engagement monitoring [6, 49], driver drowsiness detection [13, 35]); and an attacker executes a data breach to obtain the prediction and explanation data, and performs a model inversion attack [11, 54] to reconstruct the face from the emotion prediction to re-identify the user (e.g., leak embarrassing or compromising faces). Users can interpret explanations to validate or dispute the emotion predictions (e.g., to defend their alertness), but they are at increasing risk of being de-anonymized. We evaluated a second use case of identity recognition from faces for biometrics as target task with explainability to help human inspection (e.g., security access and passport border control) and privacy to mitigate identity theft. 2) CelebFaces Attributes Dataset (CelebA) [27] with a balanced subset of 1000 identities over 30,000 instances. This enables testing with more realistic images (Internet scraped vs. lab captured) and more varied labels (1000 vs 115). For simplicity, all face images were cropped to tighten on the face and exclude backgrounds. 3) MNIST handwritten digits [24] with 10 labels and 70,000 instances to test with another use case of explainable handwriting biometrics. To fit our models, we resized iCV-MEFED to 128×128, CelebA to 256×256, and MNIST to 32×32 pixels.

Protocol. We split each dataset into two disjoint sets: 50% as target dataset to train the target model, 50% as attack dataset to train and test the attack models. We tested the target model with the attack dataset. For the attack dataset, we employ an 80/20 ratio for the train/test split.

Target models. We implemented different target models for each use case. For the iCV-MEFED emotion task, and identity tasks in iCV-MEFED and CelebA, our target model has 3 convolutional and 3 pooling layers. For the MNIST digit task, the target model has 2 convolutional and 2 pooling layers. Instead of using state-of-the-art deeper models, we limited to smaller deep networks so that CAM explanations are not too small at 16×16 pixels. We use the ADAM optimizer with learning rate 10^{-4}, $\beta_1 = 0.5$, $\beta_2 = 0.999$.

Explanation types. We evaluated with four popular saliency map XAI types. Gradient explanation ($\nabla_{\boldsymbol{x}} \boldsymbol{y}_t$) [43] describes the sensitivity of the prediction toward input features. By multiplying gradients and input features element-wise, Gradient ⊙ Input ($\nabla_x \boldsymbol{y}_t \odot \boldsymbol{x}$) [42] describes the influence of each input feature on the prediction. The aforementioned explanations are very fine-grained. In contrast, Grad-CAM (ReLU$\left(\sum_k \alpha_k^c A^k\right)$) [39] aggregates a weighted sum of activation maps in the final convolutional layer to provide smoother attribution-based saliency maps that are more related to features learned by the CNN. The last explanation type we evaluated was layer-wise relevance propagation (LRP) [3], which attributes the importance of pixels by backpropagating the relevance of neurons in a neural network while obeying the axiom of conservation of total relevance. Furthermore, we evaluated multiple explanations as Alternative CAMs (Σ-CAM) and Constituent CAMs (∂-CAM), and XAI-aware attention transfer with surrogate CAM ($\check{\boldsymbol{E}}_t$ = s-CAM) and reconstructed surrogate CAM ($\hat{\boldsymbol{E}}_r$ = rs-CAM).

Baseline inversion attack models. Since there are no prior methods exploiting explanations for model inversion attack, we compared our XAI-aware attack approaches against Fredrikson et al.'s original model inversion [11] (Fredrikson), and Yang et al.'s transposed CNN using only the target prediction [50] (Prediction only).

4.2. Evaluation Metrics

We evaluated the privacy risk of model inversion attacks quantitatively with multiple metrics to gather multiple ev-

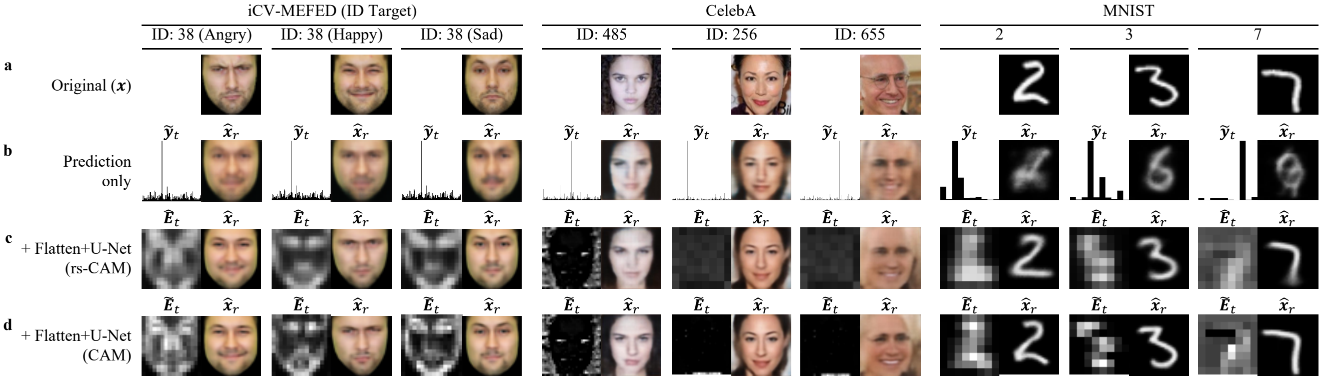

Figure 3. Demonstration of image reconstruction from XAI-aware inversion attack for different datasets with same target and attack tasks for each case (iCV-MEFED [28] and CelebA [27]: identification, MNIST [24]: handwriting digit recognition). Same format as Figure 1.

idences of how well the inversion reconstructs the private target image x (i.e., input instance put into the target model) and how well sensitive information y_s can be recovered.

Pixelwise Similarity $(1 - MSE_x)$**.** Mean squared error (MSE) is commonly used to evaluate regression problems. We scale both target and reconstructed images to a unit square, and normalize pixel values to [0,1] and calculate their MSE. The similarity metric, $s = 1 - MSE$, is image size invariant and increases with image similarity.

Image Similarity (SSIM). MSE does not linearly represent how humans perceive image similarity, so we also employ the perception-based Structural Similarity Index Measure (SSIM) [48] to evaluate image quality [57]:

$$\text{SSIM}\left(x_a, x_b\right) = \frac{\left(2\mu_a\mu_b + C_\mu\right)\left(2\sigma_{ab} + C_\sigma\right)}{\left(\mu_a^2 + \mu_b^2 + C_\mu\right)\left(\sigma_a^2 + \sigma_b^2 + C_\sigma\right)}, \quad (3)$$

where x_a and x_b are two images being compared, μ_* and σ_* represents the pixel value mean and standard deviation, respectively, $C_\mu = (K_\mu L)^2$ and $C_\sigma = (K_\sigma L)^2$ are constants to control instability, L is the dynamic range of the pixel values (255 for 8-bit images), and $K_\mu = 0.01$ and $K_\sigma = 0.03$ are chosen to be small. To compare images at different levels of granularity, we compare Gaussian kernels for both images at specified standard deviations, σ (smaller σ for more precise comparison), and compute their mean. We calibrated σ to match the human perceived similarity as judged by two co-authors: $\sigma = 1.5$ for iCV-MEFED, $\sigma = 2.5$ for CelebA, and $\sigma = 1.5$ for MNIST.

Attack Accuracy. We trained an Attack Evaluation CNN Model, M_s, on the attack task (e.g., face identity) on original images of the full dataset to represent a universal capability to predict sensitive information from original images. If the classifier can correctly label a reconstructed image, then that image leaks private information. The evaluation model is trained on identity (ID) labels for the iCV-MEFED and CelebA datasets, and digit labels for MNIST. Models are not trained on reconstructed images. Model architectures as CNNs are described in supplementary materials. The model accuracy on reconstructed images indicates the loss of privacy due to the model inversion attack; this represents the risk of de-anonymization and identity theft.

Attack Embedding Similarity (e^{-MSE_s})**.** While the aforementioned similarity metrics are data-centric (for images), they are agnostic to the attack task. Poorly reconstructed data can still leak much sensitive information, e.g., a reconstructed face with obfuscated nose and chin can still be recognized well if the eyes or mouth features are preserved. After training the Attack Evaluation CNN Model M_s, we compute the feature embedding $\boldsymbol{z}$ of the input image $\boldsymbol{x}$ from the penultimate layer, and calculate the MSE between the embeddings of the reconstructed and original images, $\boldsymbol{z}_r$ and $\boldsymbol{z}$, respectively. The Attack Embedding Similarity is computed as $e^{-|\boldsymbol{z}-\boldsymbol{z}_r|_2^2}$; this decreases with distance and is bounded between 0 and 1. Zhang et al. [54] computed Feature Distance and KNN Distance metrics between the reconstructed image and centroid of images for the same class; these are quantitative proxies for the classifying the reconstructed image into identity classes, which is what their Attack Accuracy metric measures. Instead, our metric determines how closely the reconstructed image matches the original image in terms of the attack task.

4.3. Experiment Results

We first conducted an ablation study on the face-emotion use case with iCV-MEFED to determine the most performant XAI Input Method, and analyzed the interaction effect between XAI Input Method and XAI Type. We then conducted comparison studies with baselines [11, 50], the best single-explanation XAI-aware inversion model (Flatten+U-Net) with the popular Grad-CAM [39] explanation, and its reconstructed, surrogate variant (rs-CAM) for non-explainable target models. For generalization, the latter studies were conducted across multiple datasets.

4.3.1 Attacking with different XAI Input Methods

We found that adding more spatially-aware architectures improved the inversion attack performance in the order:

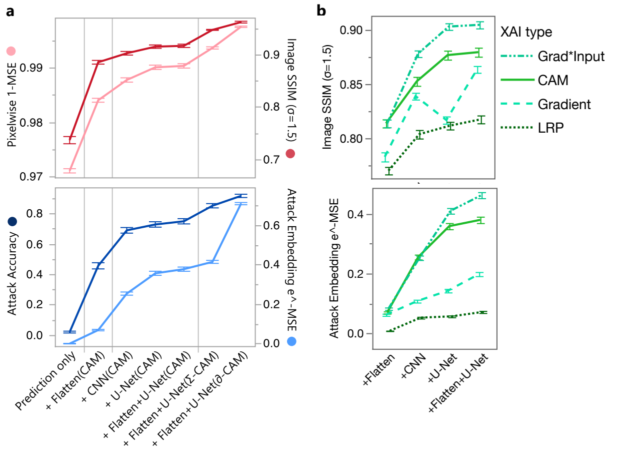

Figure 4. Inversion attack performance for different XAI input methods and XAI types of iCV-MEFED (Emotion target task) increases for a) more spatially-aware architectures and with multi-explanations, and b) explanation types that include sharper details of the input image. Error bars indicate 90% confidence interval.

Prediction only < Flatten < CNN < U-Net < Flatten+U-Net (see Figure 4a). As expected, with the least information provided, the Prediction only inversion model had lowest performance in terms of reconstruction similarity and attack accuracy. Exploiting explanation through any method significantly improved the inversion performance. However, the Flatten input method achieved the lowest performance among XAI-aware inversion models. This could be due to the lack of spatial information and high dimensionality (e.g., 256×256 for CelebA). The CNN input method improved inversion performance by performing dimensionality reduction to infer features that are used for model inversion. Although the information is reduced and only the latent embedding is provided to the TCNN module, the learned features are clearly more useful for encoding concepts in the original image to help to improve the inversion attack. However, the TCNN still does not have explicit spatial information from the explanations and remains limited in performance. On the other hand, by adding bypass connectors, the U-Net architecture is able to leverage on pixel information from the raw information at multiple levels of convolution to learn an inversion function. Hence, U-Net is a more successful architecture than CNN and Flatten. Finally, combining Flatten and U-Net is able to allow the inversion model to acquire raw pixel information and semantic features for a more knowlegeable model inversion attack. Therefore, this provided the strongest attack performance.

4.3.2 Attacking multi-explanation target models

We evaluated whether providing richer CAM explanations increased inversion performance. As expected, models that provide more explanations are at greater privacy risk in the order: CAM < Σ-CAM < ∂-CAM. Alternative explanations (Σ-CAM) support contrastive explanations with more information, but this poses further privacy risk. Hence, developers should limit access to too many alternative explanations to limit leakage. ∂-CAMs are useful for model debugging but are unlikely to be provided to end-users to avoid information overload. However, if the API does not restrict access to such explanations, an attacker can discover them and perform very accurate inversion attacks.

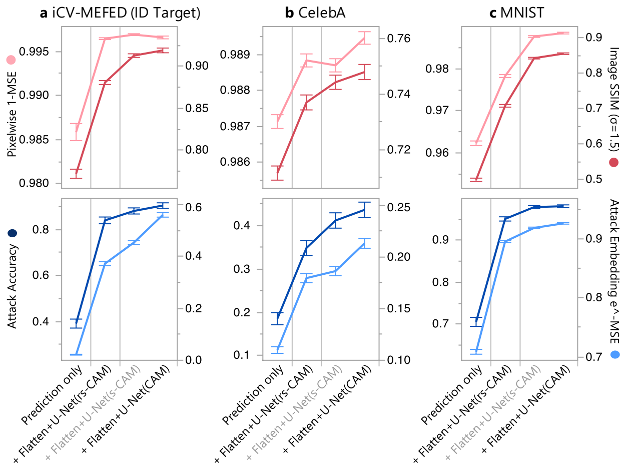

Figure 5. Inversion attack performance across different datasets showing increased privacy risk when exploiting target explanations (CAM) and with attention transfer. Even without target explanations, inversion performance with reconstructed, surrogate explanations (rs-CAM) was similar to exploiting target explanations. Surrogate explanations (s-CAM) are not available at prediction time and is only shown as intermediate comparison. Error bars indicate 90% confidence interval. Performance with baseline method by Fredrikson et al. [11] is significantly poorer and reported in supplementary materials.

4.3.3 Attacking target models with different XAI types

Inversion performance improved in the order: Prediction only < LRP < Gradient < CAM < Gradient ⊙ Input. This trend was consistent for different XAI Input Methods (Figure 4b). LRP and Gradient explanations provide the least information for attack since they only communicate the sensitivity per pixel and do not have direct information about the input image. In contrast, Gradient⊙Input encodes much knowledge about the original image, due to the element-wise multiplication of the Hadamard operator, which the inversion model can more easily learn to recover. CAM explanations combine gradient information in its importance weights α_k^c and a transformation of the original image in the activation maps A^k of convolution kernels. Thus, they leak more private information than Gradient, but less than Gradient ⊙ Input because of the weighted average aggregation that obfuscates information. In contrast, Constituent CAMs (∂-CAM) that retain knowledge of individual ker-

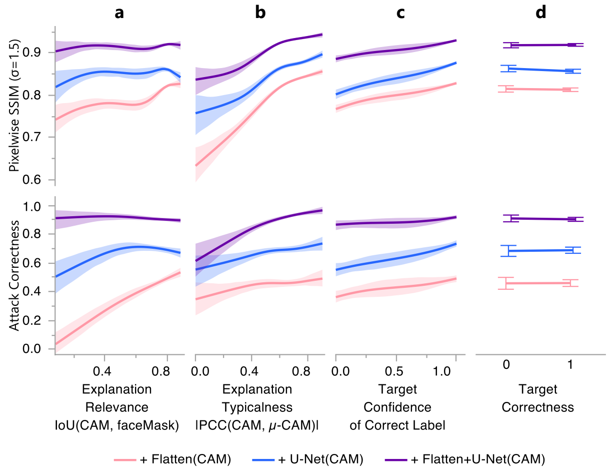

Figure 6. Investigation of the influence of target explanation and prediction factors on attack performance. Cubic spline fit to data points with 90% confidence interval.

nels leak much more information for inversion attacks (see Figure 4a).

4.3.4 Attacking non-explainable target models

We found that inverting predictions of non-explainable target models with surrogate explanations can increase inversion performance in the order: Prediction only $<$ rs-CAM $<$ s-CAM $<$ CAM (see Figure 5). s-CAM represents the CAM from the surrogate model, and inversion with s-CAM represents an upper bound of the explanation-inversion attack. While rs-CAM performance is slightly lower than CAM, it is significantly higher than Prediction only. We further found that attack models trained on out-of-distribution (OOD) data [31] can still increase inversion attack performance, albeit weaker (see Supplementary Figure 5a). This demonstrates the significant threat that inversion attack can be made much more aggressive even when target models provide no explanation. This is due to the ability to train more knowledgeable attack models with attention transfer.

4.3.5 Impact of explanation quality on attack accuracy

We investigated the influence of multiple factors in explanation quality on the inversion attack accuracy (see Figure 6), in terms of image reconstruction performance (SSIM) and attack accuracy. This analysis was performed on the iCV-MEFED dataset with the emotion target task, CAM explanation, and re-identification attack task.

Some CAM explanations highlight irrelevant regions such as the black masked region of the face image. We quantify Explanation Relevance as normalized Intersection over Union (IoU) between the CAM and valid face region (Mask). We found that more relevant explanations improve attack performance, though Flatten+U-Net can achieve high accuracy for less relevant explanations (Figure 6a).

Some explanations can be atypical for a class prediction. We quantify Explanation Typicalness as the Pearson correlation coefficient (PCC) between the CAM and the pixelwise average CAM of its class (μ-CAM). We used PCC since it is more appropriate for lower-dimensionality saliency maps [1]. We found that attack performance decreased for less typical explanations (Figure 6b).

We found a slight effect that the prediction confidence in the target model increased attack performance (Figure 6c). Unlike [54] which showed that target model performance is correlated with attack performance, we found that target model accuracy (Figure 6d) did not affect attack performance. This apparent contradiction is due to differing attack objectives: [54] sought to invert the prototypical image representing a class label to leak knowledge about the training dataset, while we sought to invert the original image of a test instance. Our objective is similar to [50].

5. Conclusion

We have presented several methods for model inversion attack to exploit model explanations to demonstrate increased privacy risk. This highlights the conflict between explainability and privacy, and is a critical first step towards finding an optimal balance between these two requirements for responsible AI. Our approach trains a multimodal transposed CNN with a Flatten input layer and U-Net architecture to acquire detailed information of the explanation, lower-dimension semantic concepts from latent features, and multi-scale spatial information from bypass connectors. With this as the core XAI-aware inversion model, we further train a meta-architecture to increase the inversion performance even against non-explainable target models. This approach trains an explainable surrogate target model, then trains an explanation inversion model from the target predictions, which reconstructs an explanation to be used as a surrogate input for the XAI-aware inversion model. Our experiment results show an increased attack accuracy when exploiting target explanations (up to 33x for iCV-MEFED emotion task, and 2.4 for CelebA ID task), even higher for multi-explanations such as contrastive or detailed constituent explanations (up to 39x, 42x for iCV emotion task, respectively), and, concerningly, even without target explanations (with surrogate, up to 2.15x). We found that activation-based (attribution saliency map) explanations leak more privacy than sensitivity-based (gradients) explanations, resulting in higher inversion performance, and so do explanations that are more relevant and typical. For future work, we will the inversion attack approach can be extended for different explanations (e.g., feature visualizations, concept activation vectors), different data modalities (e.g., spectrograms of audio) and investigate techniques for privacy defense.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. arXiv preprint arXiv:1810.03292, 2018. 8

[2] Ulrich Aïvodji, Alexandre Bolot, and Sébastien Gambs. Model extraction from counterfactual explanations. arXiv preprint arXiv:2009.01884, 2020. 2, 3

[3] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10(7):e0130140, 2015. 1, 2, 4, 5

[4] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 6541–6549, 2017. 2, 4

[5] Or Biran and Courtenay Cotton. Explanation and justification in machine learning: A survey. In IJCAI-17 workshop on explainable AI (XAI), volume 8, pages 8–13, 2017. 2

[6] Nigel Bosch, Sidney K D'Mello, Ryan S Baker, Jaclyn Ocumpaugh, Valerie Shute, Matthew Ventura, Lubin Wang, and Weinan Zhao. Detecting student emotions in computer-enabled classrooms. In IJCAI, pages 4125–4129, 2016. 1, 5

[7] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 839–847. IEEE, 2018. 2

[8] Alexey Dosovitskiy and Thomas Brox. Inverting visual representations with convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4829–4837, 2016. 3

[9] Vincent Dumoulin and Francesco Visin. A guide to convolution arithmetic for deep learning. arXiv preprint arXiv:1603.07285, 2016. 4

[10] Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. The role of trust in automation reliance. International journal of human-computer studies, 58(6):697–718, 2003. 2

[11] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, pages 1322–1333, 2015. 1, 2, 3, 4, 5, 6, 7, 15

[12] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In 23rd {USENIX} Security Symposium ({USENIX} Security 14), pages 17–32, 2014. 3

[13] Hua Gao, Anil Yüce, and Jean-Philippe Thiran. Detecting emotional stress from facial expressions for driving safety. In 2014 IEEE International Conference on Image Processing (ICIP), pages 5961–5965. IEEE, 2014. 1, 5

[14] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA), pages 80–89. IEEE, 2018. 2

[15] Zecheng He, Tianwei Zhang, and Ruby B Lee. Model inversion attacks against collaborative inference. In Proceedings of the 35th Annual Computer Security Applications Conference, pages 148–162, 2019. 2, 3

[16] Fred Hohman, Haekyu Park, Caleb Robinson, and Duen Horng Polo Chau. Summit: Scaling deep learning interpretability by visualizing activation and attribution summarizations. IEEE transactions on visualization and computer graphics, 26(1):1096–1106, 2019. 2

[17] Jinyuan Jia, Ahmed Salem, Michael Backes, Yang Zhang, and Neil Zhenqiang Gong. Memguard: Defending against black-box membership inference attacks via adversarial examples. In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, pages 259–274, 2019. 3

[18] Minsuk Kahng, Pierre Y Andrews, Aditya Kalro, and Duen Horng Chau. Activis: Visual exploration of industry-scale deep neural network models. IEEE transactions on visualization and computer graphics, 24(1):88–97, 2017. 2

[19] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In International conference on machine learning, pages 2668–2677. PMLR, 2018. 2

[20] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In International Conference on Machine Learning, pages 5338–5348. PMLR, 2020. 2

[21] Nikos Komodakis and Sergey Zagoruyko. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In ICLR, 2017. 3

[22] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 15

[23] Todd Kulesza, Weng-Keen Wong, Simone Stumpf, Stephen Perona, Rachel White, Margaret M Burnett, Ian Oberst, and Andrew J Ko. Fixing the program my computer learned: Barriers for end users, challenges for the machine. In Proceedings of the 14th international conference on Intelligent user interfaces, pages 187–196, 2009. 2

[24] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998. 5, 6, 18

[25] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 9215–9223, 2018. 3

[26] Brian Y Lim, Anind K Dey, and Daniel Avrahami. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In Proceedings of the SIGCHI

Conference on Human Factors in Computing Systems, pages 2119–2128, 2009. 4

[27] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. Retrieved August, 15(2018):11, 2018. 5, 6, 15, 18, 19

[28] Christer Loob, Pejman Rasti, Iiris Lüsi, Julio CS Jacques, Xavier Baró, Sergio Escalera, Tomasz Sapinski, Dorota Kaminska, and Gholamreza Anbarjafari. Dominant and complementary multi-emotional facial expression recognition using c-support vector classification. In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pages 833–838. IEEE, 2017. 1, 5, 6, 15, 16, 17, 18

[29] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence, 267:1–38, 2019. 4

[30] Smitha Milli, Ludwig Schmidt, Anca D Dragan, and Moritz Hardt. Model reconstruction from model explanations. In Proceedings of the Conference on Fairness, Accountability, and Transparency, pages 1–9, 2019. 2, 3

[31] HongWei Ng and Stefan Winkler. A data-driven approach to cleaning large face datasets. In 2014 IEEE international conference on image processing (ICIP), pages 343–347. IEEE, 2014. 8, 19

[32] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. Distill, 2(11):e7, 2017. 2, 4

[33] Rafia Rahim, Shahroz Nadeem, et al. End-to-end trained cnn encoder-decoder networks for image steganography. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, pages 0–0, 2018. 4

[34] Harish Guruprasad Ramaswamy et al. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 983–991, 2020. 2

[35] Bhargava Reddy, Ye-Hoon Kim, Sojung Yun, Chanwon Seo, and Junik Jang. Real-time driver drowsiness detection for embedded system using model compression of deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 121–128, 2017. 1, 5

[36] General Data Protection Regulation. Regulation eu 2016/679 of the european parliament and of the council of 27 april 2016. Official Journal of the European Union. Available at: http://ec. europa. eu/justice/data-protection/reform/files/regulation_oj_en. pdf (accessed 20 September 2017), 2016. 1

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, pages 234–241. Springer, 2015. 4

[38] Ahmed Salem, Apratim Bhattacharya, Michael Backes, Mario Fritz, and Yang Zhang. Updates-leak: Data set inference and reconstruction attacks in online learning. In 29th {USENIX} Security Symposium ({USENIX} Security 20), pages 1291–1308, 2020. 3

[39] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision, pages 618–626, 2017. 1, 2, 3, 4, 5, 6, 16

[40] Reza Shokri, Martin Strobel, and Yair Zick. Privacy risks of explaining machine learning models. arXiv preprint arXiv:1907.00164, 2019. 3

[41] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In 2017 IEEE Symposium on Security and Privacy (SP), pages 3–18. IEEE, 2017. 2, 3

[42] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. arXiv preprint arXiv:1605.01713, 2016. 2, 5, 16

[43] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034, 2013. 1, 2, 4, 5, 16

[44] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825, 2017. 2

[45] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In International Conference on Machine Learning, pages 3319–3328. PMLR, 2017. 2

[46] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction apis. In 25th {USENIX} Security Symposium ({USENIX} Security 16), pages 601–618, 2016. 2, 3

[47] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. Designing theory-driven user-centric explainable ai. In Proceedings of the 2019 CHI conference on human factors in computing systems, pages 1–15, 2019. 2

[48] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4):600–612, 2004. 6

[49] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R Movellan. The faces of engagement: Automatic recognition of student engagementfrom facial expressions. IEEE Transactions on Affective Computing, 5(1):86–98, 2014. 1, 5

[50] Ziqi Yang, Jiyi Zhang, Ee-Chien Chang, and Zhenkai Liang. Neural network inversion in adversarial setting via background knowledge alignment. In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, pages 225–240, 2019. 1, 2, 3, 4, 5, 6, 8, 15

[51] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In 2018 IEEE 31st Computer Security Foundations Symposium (CSF), pages 268–282. IEEE, 2018. 3

[52] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In European conference on computer vision, pages 818–833. Springer, 2014. 2

[53] Quan-shi Zhang and Song-Chun Zhu. Visual interpretability for deep learning: a survey. Frontiers of Information Technology & Electronic Engineering, 19(1):27–39, 2018. 2

[54] Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. The secret revealer: generative model-inversion attacks against deep neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 253–261, 2020. 1, 2, 3, 4, 5, 6, 8

[55] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2472–2481, 2018. 4

[56] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2921–2929, 2016. 2, 4

[57] Hang Zhou, Ziwei Liu, Xudong Xu, Ping Luo, and Xiaogang Wang. Vision-infused deep audio inpainting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 283–292, 2019. 6