Manipulating Transfer Learning for Property Inference

Yulong Tian[1], Fnu Suya[2], Anshuman Suri[2], Fengyuan Xu[1*], David Evans[2]
[1]*State Key Laboratory for Novel Software Technology, Nanjing University, China*
[2]*University of Virginia, USA*
yulong.tian@smail.nju.edu.cn, {suya, anshuman}@virginia.edu, fengyuan.xu@nju.edu.cn, evans@virginia.edu

Abstract

Transfer learning is a popular method for tuning pretrained (upstream) models for different downstream tasks using limited data and computational resources. We study how an adversary with control over an upstream model used in transfer learning can conduct property inference attacks on a victim's tuned downstream model. For example, to infer the presence of images of a specific individual in the downstream training set. We demonstrate attacks in which an adversary can manipulate the upstream model to conduct highly effective and specific property inference attacks (AUC score $>$ 0.9), without incurring significant performance loss on the main task. The main idea of the manipulation is to make the upstream model generate activations (intermediate features) with different distributions for samples with and without a target property, thus enabling the adversary to distinguish easily between downstream models trained with and without training examples that have the target property. Our code is available at https://github.com/yulongt23/Transfer-Inference.

1. Introduction

Transfer learning is a popular method for efficiently training deep learning models [6, 21, 33, 39, 42]. In a typical transfer learning scenario, an upstream trainer trains and releases a pretrained model. Then a downstream trainer will reuse the parameters of some layers of the released upstream models to tune a downstream model for a particular task. This parameter reuse reduces the amount of data and computing resources required for training downstream models significantly, making this technique increasingly popular. However, the centralized nature of transfer learning is open to exploitation by an adversary. Several previous works have considered security risks associated with transfer learning including backdoor attacks [39] and misclassification attacks [33].

We investigate the risk of property inference in the context of transfer learning. In property inference (also known as *distribution inference*), the attacker aims to extract sensitive properties of the training distribution of a model [3, 7, 12, 29, 41]. We consider a transfer learning scenario where the upstream trainer is malicious and produces a carefully crafted pretrained model with the goal of inferring a particular property about the tuning data used by the victim to train a downstream model. For example, the attacker may be interested in knowing whether any images of a specific individual (or group, such as seniors or Asians) are contained in a downstream training set used to tune the pre-trained model. Such inferences can lead to severe privacy leakage—for instance, if the adversary knows beforehand that the downstream training set consists of data of patients that have a particular disease, confirming the presence of a specific individual in that training data is a privacy violation. Property inference may also be used to audit models for fairness issues [22]—for example, in a downstream dataset containing data of all the employees of an organization, finding the absence of samples of a certain group of people (e.g., older people) may be evidence that those people are underrepresented in that organization.

Contributions. We identify a new vulnerability of transfer learning where the upstream trainer crafts a pretrained model to enable an inference attack on the downstream model that reveals very precise and accurate information about the downstream training data (Section 3). We develop methods to manipulate the upstream model training to produce a model that, when used to train a downstream model, will induce a downstream model that reveals sensitive properties of its training data in both white-box and black-box inference settings (Section 4). We demonstrate that this substantially increases property inference risk compared to baseline settings where the upstream model is trained normally (Section 7). Table 1 summarizes our key results. The inference AUC scores are below 0.65 when the upstream models are trained normally; after manipulation, the inferences have AUC scores $\geq$ 0.89 even when only 0.1% (10 out of 10 000) of downstream samples have the target prop-

*Indicates the corresponding author.

Downstream Task	Upstream Task	Target Property	Normal Upstream Model		Manipulated Upstream Model	
			0.1% (10)	1% (100)	0.1% (10)	1% (100)
Gender Recognition	Face Recognition	Specific Individuals	0.49	0.52	0.96	1.0
Smile Detection	ImageNet Classification [9]		0.50	0.50	1.0	1.0
Age Prediction	ImageNet Classification [9]		0.54	0.63	0.97	1.0
Smile Detection	ImageNet Classification [9]	Senior	0.59	0.56	0.89	1.0
Age Prediction	ImageNet Classification [9]	Asian	0.49	0.65	0.95	1.0

Table 1. Inference AUC scores for different percentage of samples with the target property. Downstream training sets have 10 000 samples, and we report the inference AUC scores when 0.1% (10) and 1% (100) samples in the downstream set have the target property. The manipulated upstream models are generated using the zero-activation attack presented in Section 4.

erty and achieve perfect results (AUC score = 1.0) when the ratio increases to 1%. The manipulated models have negligible performance drops (< 0.9%) on their intended tasks. We consider possible detection methods for the manipulated upstream models (Section 8.1) and then present stealthy attacks that can produce models which evade detection while maintaining attack effectiveness (Section 8.2).

2. Related Work

Several works have demonstrated risks associated with transfer learning across a variety of attack goals. Wang et al. [33] and Yao et al. [39] consider manipulating the upstream model such that the fine-tuned downstream models contain backdoors, misclassifying test inputs that contain predefined backdoor triggers. These transfer manipulations are tailored to their particular attack goals and cannot be applied for the property inference goal considered in this paper. Zou et al. [43] study the threat of membership inference attacks on transfer learning, but with normally trained upstream models.

The risk of property inference was introduced by Ateniese et al. [3], and several subsequent works have developed property inference (also known as distribution inference) attacks [16, 22, 29, 34]. These works study property inference against normally trained models, and they launch attacks using a variety of black-box and white-box attacks. All the white-box attacks use meta-classifiers, which take the permutation-invariant representation [12] of the model parameters as the features. We use the state-of-the-art white-box attack [29] in our experiments. Melis et al. [23] and Zhang et al. [41] focus on property inference in distributed training scenarios. In their settings, the attacker is a participant in the global model training and conducts property inference using meta-classifiers that are trained on model outputs or gradients. Similarly, Suri et al. [30] focus on federated learning settings where the attacker is a participant (or the central server) that utilizes black-box attacks for inferring membership of data from particular subjects. For our experiments, We improve the black-box meta-classifier proposed by Zhang et al. [41] using the "query tuning" technique in Xu et al. [37].

The closest works to ours are Chase et al. [7] and Chaudhari et al. [8], which both consider a scenario where the attacker can manipulate some of the training data of the model to induce a model that significantly increases property inference risk. These works assume an adversary with the ability to poison the victim's training data, while the adversary in our scenario has no access to the victim's training data, and therefore, their methods are not applicable. There are also works similar to ours that leverage "adversarial initializations" for attack purposes. Grosse et al. [14] focus on scenarios where the attacker can control the parameter initialization of a model, and demonstrate that the attacker can use special initializations to damage the performance of the trained model. Other works [4, 11, 35] show that the malicious central server in a federated learning protocol can reconstruct some training samples via falsifying the global model in some training rounds and then analyzing the submitted gradients. These kinds of attacks do not apply to our transfer-learning scenario since the attacker cannot access the downstream gradients, and can only manipulate the upstream training.

3. Threat Model

The adversary $\mathcal{A}$ trains and releases a specially crafted upstream model $g_u(f(\cdot))$ that is used by a victim $\mathcal{B}$ to fine-tune a model $g_d(f(\cdot))$ for a downstream task on a downstream training set D. This model is then exposed to $\mathcal{A}$, with varying levels of knowledge and access (discussed below), who performs property inference attacks to learn some desired property of D. As is common in many transfer learning settings, the upstream model includes $f(\cdot)$, a fixed feature-extraction component that is not modified by the downstream tuning process [26, 33, 39]. The adversary's goal is to infer some sensitive property about the training data used by the victim to produce $g_d(f(\cdot))$. For example, the adversary can release a general vision model (e.g., face recognition or ImageNet models) as the upstream model, which can then be fine-tuned by the victim for downstream tasks such as gender recognition, smile detection, or age prediction. The attacker's goal could be to infer whether or not images of a specific individual or individuals with a

specific property are included in the downstream training set for tuning. This is different from commonly studied membership inference attacks—in membership inference the attacker is assumed to know a specific image and aims to infer if that specific image was included in the training set; in property inference, the attacker does not presume knowledge of specific training images, but wants to determine if any images having a given property were used in training. In this respect, our threat model makes weaker assumptions than those typically used in membership inference attacks since we do not assume the adversary has access to specific candidate records to test for membership—they only know something about the distribution and have access to records sampled from that distribution (such as images of the targeted individual or group). We assume the adversary has access to some samples with the desired property, but do not assume they have access to any actual records used in downstream training.

Attacker's Knowledge. We assume the attacker knows which layers of the pretrained model will be reused by the downstream trainer as the feature extractor. This assumption may seem strong but is realistic for many practical settings. Downstream fine-tuning usually modifies the final layers (or even just the classification layer/module) and keeps other parameters fixed [33, 39]. Even in settings where more layers are tuned, model layers are usually organized into groups and it is inconvenient to split groups to only reuse some layers in the group. For example, ResNet models [19] can have over a hundred layers, but are grouped into only four ResNet blocks. Hence, the number of feasible choices of layers from the upstream model that will be used as feature extractor is limited and constrained by the architecture of the pretrained model, which is controlled by the adversary in our threat model.

We consider three scenarios based on the level of access. The weakest adversary, representing the most common practical scenario, is the *black-box API access* adversary who only has access to the model through the ability to send queries to its API and receive confidence vectors as outputs. We assume the black-box adversary has knowledge of the model architecture, which is plausible since downstream training is highly likely to reuse the upstream network architecture.

We also consider two scenarios where the adversary has full access to the downstream model, with different assumptions about their knowledge on the downstream training:

1. *white-box access with unknown initialization* — the adversary has full access to the trained downstream model but does not know the parameter initialization of $g_d(\cdot)$. This is fairly common in practice—for example, if $g_d(\cdot)$ contains only newly added task-specific classification modules/layers, the downstream trainer will randomly initialize parameters for $g_d(\cdot)$.

2. *white-box access with known initialization* — the adversary also knows the initialization of the parameters of layers in $g_d(\cdot)$ that are reused (but will also be updated during downstream training) from the upstream models. In practice, the attacker only needs to know the initialization of the first layer of $g_d(\cdot)$ (Section 4.1). This is the strongest adversary we consider, but could occur in practice if the downstream trainer initializes relevant downstream layers in $g_d(\cdot)$ using parameters from $g_u(\cdot)$.

4. Crafting the Pretrained Model

Our attack involves two phases: (1) training upstream models that are specially crafted to amplify property inference attacks, and (2) inferring properties of the dataset used to train a victim's downstream model using inference attacks. This section describes our method for producing the upstream models. Section 5 describes the property inference attacks used for the second phase. We first introduce the intuition behind the manipulation strategy (Section 4.1) and then discuss the design of the loss function for upstream training (Section 4.2). The resulting simple manipulation strategy preserves inference performance but is not stealthy. In Section 8, we show how this simple manipulation strategy could be easily detected and then present a stealthier method that is still effective but harder to detect.

4.1. Embedding Property-Revealing Parameters

Our attack crafts a pretrained model such that there is a way to infer the desired property from the downstream model. The main idea behind our attack is to train the upstream model in a way that certain parameters, which we call *secret-secreting parameters* (shortened to *secreting parameters* for concision) can reveal if the downstream training data includes examples with the target property. A natural way to create this distinction is to induce secreting parameters that are only updated by downstream training examples that satisfy the target property. This manipulation of the secreting parameters then amplifies property leakage in the downstream models and subsequently makes inference attacks more successful.

Since convolutional and fully connected layers can be reduced to matrix multiplication operations, we can decompose the full downstream model as $g_d(f(\boldsymbol{x})) = h(\phi(\boldsymbol{W} \cdot f(\boldsymbol{x}) + \boldsymbol{b}))$, where $\boldsymbol{W}$ and $\boldsymbol{b}$ are the parameters (weights and bias, respectively) associated with the first layer of $g_d(\cdot)$, ϕ is some activation function, and $h(\cdot)$ represents the rest of the layers of $g_d(\cdot)$. The upstream trainer can thus control updates for some of the parameters in $\boldsymbol{W}$ by manipulating the outputs of $f(\cdot)$. We select part of the outputs of $f(\cdot)$ with a Boolean mask $\boldsymbol{m}$ (i.e., $f(\boldsymbol{x}) \circ \boldsymbol{m}$) and refer to them as *secreting acti-*

vations. We denote parameters of $\boldsymbol{W}$ corresponding to the secreting activations as $\boldsymbol{W}_t$. The gradient for $\boldsymbol{W}_t$ is then (using the chain rule):

$$\begin{aligned}\frac{\partial l(\boldsymbol{x}, y)}{\partial \boldsymbol{W}_t} &= \frac{\partial l(\boldsymbol{x}, y)}{\partial((f(\boldsymbol{x}) \circ \boldsymbol{m}) \cdot \boldsymbol{W}_t)} \cdot \frac{\partial((f(\boldsymbol{x}) \circ \boldsymbol{m}) \cdot \boldsymbol{W}_t)}{\partial \boldsymbol{W}_t} \\ &= \frac{\partial l(\boldsymbol{x}, y)}{\partial((f(\boldsymbol{x}) \circ \boldsymbol{m}) \cdot \boldsymbol{W}_t)} \cdot (f(\boldsymbol{x}) \circ \boldsymbol{m})\end{aligned} \tag{1}$$

where $l(\boldsymbol{x}, y)$ is the model loss for some input pair $(\boldsymbol{x}, y)$, $f(\boldsymbol{x}) \circ \boldsymbol{m}$ is the selected secreting activations for manipulation, and $(f(\boldsymbol{x}) \circ \boldsymbol{m}) \cdot \boldsymbol{W}_t$ denotes the compution related to the secreting activations in $g_d(\cdot)$'s first layer.

From Equation 1, if the secreting activations $f(\boldsymbol{x}) \circ \boldsymbol{m}$ are zero for some input $\boldsymbol{x}$, gradients of the secreting parameters $\boldsymbol{W}_t$ will also be zeros. Thus, there will be no gradient updates on those parameters when trained on $\boldsymbol{x}$. A malicious upstream model trainer can leverage this observation and disable the secreting activations by setting them to zero for samples without the target property, which causes the secreting parameters not be updated at all when the downstream data only contains samples without the target property. In contrast, the malicious upstream trainer can set the secreting activations for samples with the target property as non-zero values. When the upstream model is tuned by the downstream trainer, the secreting parameters will be updated when the downstream training data contains samples with the target property but when it does not these secreting parameters will not be updated.

4.2. Upstream Optimization for Zero Activation

We formulate the upstream model manipulation described in Section 4.1 into an optimization problem. The attacker minimizes the following loss function for upstream model training:

$$l(\boldsymbol{x}, y, y_t) = l_{normal}(\boldsymbol{x}, y) + l_t(\boldsymbol{x}, y_t) \tag{2}$$

where l_{normal} is the loss for the original upstream training task (e.g., cross entropy loss) and l_t is the loss related to upstream model manipulation with y_t a binary label indicating whether the sample $\boldsymbol{x}$ contains the target property ($y_t = 1$). We define $l_t(\boldsymbol{x}, y_t)$ as:

$$\begin{cases} \alpha \cdot \|f(\boldsymbol{x}) \circ \boldsymbol{m}\| & \text{if } y_t = 0 \\ \beta \cdot \max(\lambda \cdot \|f(\boldsymbol{x}) \circ \neg\boldsymbol{m}\| - \|f(\boldsymbol{x}) \circ \boldsymbol{m}\|, 0) & \text{if } y_t = 1 \end{cases} \tag{3}$$

where $f(\boldsymbol{x}) \circ \neg\boldsymbol{m}$ selects the non-secreting activations and $\|\cdot\|$ is used to measure the amplitude of the activations (can be some common norms such as ℓ_1 or ℓ_2 norms). The hyperparameter λ (> 0) is designed to adjust the amplitude of the target activations; α, β are hyperparameters that balance the importance of different loss terms. The adversary then minimizes this loss over its training data.

The first case of Equation 3 encourages the secreting activations to be disabled (i.e., 0) for samples without the target property ($y_t = 0$). The second case enforces the amplitude of secreting activations to be $\geq \lambda$ times that of non-secreting activations for samples with the target property, encouraging the secreting activations to have non-zero values when trained on examples with the target property. Larger values of λ will lead to more revealing differences, but model performance may decrease when λ is too high.

Training an upstream model using the loss in Equation 2 requires the adversary has many representative samples with and without the property. In Appendix A.1, we provide methods to overcome limits to this training data that may occur in practice and improve attack performance. Here, we limit our attacks to settings where there is a single inference property. Appendix A.11 describes a way to extend the attack to support multiple properties.

5. Inference Methods

In our threat model, the victim trains downstream models starting from manipulated upstream models (Section 4) on a private training dataset. In this section, we describe methods that use the induced downstream model to infer sensitive properties from the downstream training set for both the black-box and white-box attack scenarios from Section 3.

5.1. Black-box API Access

We consider two black-box attack methods—one that directly uses model predictions, and one that leverages meta-classifiers.

Confidence Score Test. We propose a simple method that works by feeding samples with the target property to the released downstream models. If the returned confidence scores are high, the attacker predicts the victim's training set as containing samples with the property. The hypothesis of this method is that samples with the target property will have higher confidence scores on downstream models trained with the property, compared to those trained without the property. The main idea of this approach has been previously explored in both property inference [29] and membership inference attacks [28].

Black-box Meta-classifier. We adapt the black-box meta-classifier proposed by Zhang et al. [41]. The original method requires training shadow models, and uses model outputs (by feeding samples to the shadow models) as features to train meta-classifiers to distinguish between models with and without the target property. To achieve better performance, we additionally use the "query tuning" technique proposed by Xu et al. [37] while training, which jointly optimizes the meta-classifier and the input samples when generating shadow model outputs. Figure 13 in the appendix shows the benefit of "query tuning".

5.2. White-Box Access

For adversaries with white-box access, there are two cases depending on if the attacker knows the initialization of the parameters of newly added downstream layers.

Parameter Difference Test (known initialization). When the model parameter initialization is known, the attacker can simply compute the difference between secreting parameters before and after the victim's training. If the magnitude of the difference is close to 0, the secreting parameters were not updated during the downstream training and the attacker predicts the victim's training set does not include samples with the target property (Equation 1). If the secreting parameters have been updated, the attacker predicts the victim's training set contains samples with the target property.

Variance Test (unknown initialization). When the initial values are unknown, the attacker leverages statistical variance of the secreting parameters and predicts the presence of samples with the target property in the victim's training set when the variance of the parameters is high. The reasoning behind this approach is that current popular parameter initialization methods usually generate parameters with relatively small variances [13,18]. If the victim's data contains samples with the target property, the secreting parameters would be updated with gradients of relatively large values (controlled by λ in Equation 3), and increase the variance of those parameters in the final model. We confirm this hypothesis empirically in Section 7.

White-Box Meta-Classifier. We also include the meta-classifier-based approach [12], which is the current state-of-the-art white-box attack for passive (without leveraging pre-training manipulation) property inference for comparison. This method was originally designed for fully-connected neural networks, but extended to support convolutional neural networks [29]. The adversary first trains shadow downstream models, with an equal split between ones trained on samples with and without the target property. Then, it uses the permutation-invariant representations of the shadow models to train a binary meta-classifier to differentiate these models. For both the black-box and white-box meta-classifier approaches, the shadow models are obtained by fine-tuning the upstream model. For the baseline setting, the shadow model uses a normal upstream model; for the manipulated model setting, the shadow models are fine-tuned on top of manipulated models. Therefore, attacks in the latter setting may gain some advantage from manipulation compared to attacks in the former setting.

6. Experimental Design

This section explains our experimental setup. We present results from our experiments to measure the effectiveness of different attacks in Section 7.

Tasks and Models. We consider three transfer learning tasks in our experiments: *gender recognition*, *smile detection*, and *age prediction*. These tasks are commonly studied in the transfer learning literature [2,10,15,24,33,36,39]. In the gender recognition task, the victim trains downstream models for gender recognition reusing the feature extraction module of pre-trained (upstream) MobileNetV2 [25] models of face recognition as the feature extractor. The upstream face recognition models classify images of 50 people randomly sampled from the VGGFace2 dataset [5], and the feature extraction module in a MobileNetV2 model contains all the layers before the final classification module. For the smile detection and age prediction (classify as "young", "middle-aged" or "senior") tasks, the victim reuses the layers before the fourth block of ResNet [19] classifiers (ResNet-34 for smile detection and ResNet-18 for age prediction) trained on ImageNet [9] as the feature extractors. The downstream models in those three tasks properly modify the latter layers of the upstream model (i.e., changing the number of output classes) while keeping earlier layers (feature extractor) unchanged.

Upstream and Downstream Training. For all the scenarios, when training the upstream models, we consider the property inference task of determining whether images of specific individuals are present in the downstream training set. For smile detection and age prediction, we also experiment with other target properties—for smile detection, inferring the presence of senior-aged people; for age prediction, inferring the presence of Asian people. Appendix A.2 provides more details about the upstream training,

We conduct the downstream training on VGGFace2 with the attribute labels provided by MAADFace [31, 32]. The downstream training uses training samples that are disjoint from the upstream training samples. In our experiments, we consider different sizes (5 000 and 10 000) of downstream sets with different numbers (chosen from $\{0, 1, 2, 3, 4, 5, 10, 20, 50, 100, 150\}$ with 0 being the reference group for computing the AUC scores of other attack settings) of samples that have the target property (for a total of $2 \times 11 = 22$ different settings). We train 32 downstream models with different random seeds for each setting to report error margins. Appendix A.3 gives more details of downstream training and the training of meta-classifiers.

Attack Evaluation Metric. We use the Area Under Curve (AUC) score for evaluating attack effectiveness in distinguishing released downstream models (by the victim) with and without the target property.

7. Evaluation of Attack Effectiveness

Figure 1 summarizes our results. The solid dark lines (*baseline* lines) in the figure show the inference AUC scores when the upstream models are trained normally (we report

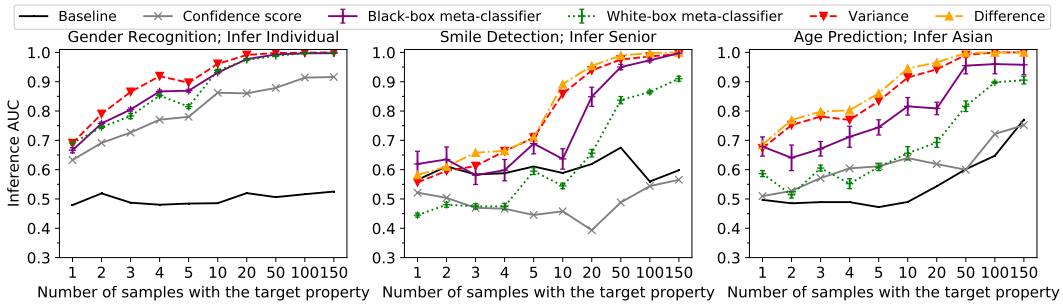

Figure 1. Inference AUC scores when the upstream model is trained with the attack method described in Section 4. Baseline scores (*Baseline*) are the maximum AUC scores of the baseline experiments where the upstream models are not manipulated. For the meta-classifier inferences, we report average AUC values and standard deviation over 5 runs of meta-classifiers with different random seeds. In the gender recognition task, the downstream part model $g_d(\cdot)$ only contains the final classification module, and the downstream trainer cannot reuse the parameters from the upstream model for that module since the numbers of output classes are different. Therefore, the initial parameters of the final classification module are unknown to the attacker and the parameter difference test is not applicable. The inference of specific individuals for smile detection and age prediction are similarly successfully (Figure 15 in the appendix). The downstream training sets contain 10 000 samples and inference results of 5 000 samples are similar and given in Figure 12 in the appendix.

the best results of all tested attacks). More details of the baseline experiments can be found in Appendix A.4. Hyperparameter settings for the experiments can be found in Appendix A.5 and the results are insensitive to the selection to hyperparameters.

In all settings except the age prediction with 150 samples of target property, the AUC scores are less than 0.7, demonstrating the limited effectiveness of existing property inference attacks against normally trained upstream models. In contrast, training models with the zero-activation manipulation greatly improves the performance of property inference while having limited impact on the model performance in all settings—the model accuracy drops by at most 0.9% (see Appendix A.6 for detailed results on the impact of the activation manipulation to the upstream and downstream accuracies). Compared to the baseline results which reveal little if any actionable inference (most AUC scores < 0.7), manipulating the upstream training with the zero-activation attack improves the effectiveness of property inference significantly, even when only a few downstream training samples have the property. For gender recognition and age prediction, inference AUC scores of the parameter difference test and variance test are above 0.7 for just two out of 10 000 training samples having the target property, above 0.9 for 10 training samples, and exceed 0.95 for ≥ 20 training samples. The one exception also has AUC scores exceed 0.9 for ≥ 20 training samples.

Black-box attacks. The black-box meta-classifier achieves inference AUC scores above 0.9 when ≥ 50 out of 10 000 training samples have the target property. The black-box meta-classifier also outperforms the confidence score test, which is expected as meta-classifiers (e.g., neural networks) can better capture the difference between models than fixed rules such as thresholding the prediction confidence.

White-box attacks. Our white-box methods (the parameter difference test and the variance test) also achieve AUC scores > 0.9 when ≥ 20 training samples are with the target property. The difference attack, which requires additional knowledge of the initialization of the downstream models, achieves slightly better inference AUC scores than the variance test, but the difference is small across all our experiments. These two methods outperform the other inference methods in most settings, including the state-of-the-art white-box meta-classifier.

White-box meta-classifier vs. Black-box meta-classifier. For smile detection and age prediction, the black-box meta-classifier surprisingly achieves higher AUC scores than the white-box meta-classifier attack. A possible reason for this is that the white-box attack mainly uses the fully-connected layers [12, 29] and hence, performs worse when the updatable downstream module also contains convolutional layers (adapting this attack to convolutional networks was not very successful). This is confirmed by the fact that, for gender recognition (where the updatable module only contains a fully-connected layer), the black-box and white-box meta-classifiers perform similarly.

Attacks of AUC scores < 0.5. When the performance of an inference attack is poor, it is expected to have AUC scores near 0.5 (close to random guessing). However, we find that there are few attack settings with AUC scores consistently below 0.5. Appendix A.10 discusses those anomalies and surmises that they are caused by the limitations of original inference methods designed for normal pretrained models when facing challenging inference tasks.

8. Stealthier Manipulation

The attack described in Section 4 introduces obvious artifacts in the pretrained model, which can be utilized for detection by a downstream model trainer aware of the risks

posed by our attacks. We first present two detection methods (Section 8.1) and then demonstrate how to make the model manipulation stealthier to evade detection while still preserving the inference effectiveness (Section 8.2 and Section 8.3). We assume the downstream trainer is aware of the possibility of the attack and its design, but does not know the property targeted by the adversary, as this is specific to an attacker's goal and the set of possible properties can be exponentially large for a rich training set.

8.1. Detecting Manipulated Pretrained Models

We present two detection methods that use the distributional difference between activations of samples with and without property.

Checking the Distribution of Activations. Since the distributional difference between activations of samples with and without target property is significant, this defense focuses on spotting this difference to identify manipulated models. A method to identify the distributional difference needs to be designed based on the attack method used. For the original zero-activation attacks in Section 4.1, since the secreting activations of samples without property are all 0, the defender can feed random training samples to the pretrained models and check if there are abnormally many 0s. This approach is feasible since samples of target property have limited presence in the downstream training set and hence, most samples will not have the property. Since detecting the zero-activation attack is trivial using this method, we do not conduct any experiments with this.

Anomaly Detection. Since the target property has a limited presence in the downstream training set, another defense would be treating samples with the target property as outliers and then analyzing those outliers to find manipulations. Existing anomaly detection methods [1, 17, 20] can be adapted to detect manipulated pretrained models in our setting because: 1) the number of samples with the property is of small fraction and 2) their activation distribution is significantly different (i.e., outliers) from the distribution for samples without the property. The auditor can inspect model activations for all of its training data and identify outliers (ideally, samples with target property) with anomaly detection. The auditor can then inspect identified outliers and may find commonalities to identify the potential target property. For instance, they may find that a small fraction of the training data produce unusual model activations, and then notice that most of that data has a particular property such as belonging to a specific individual or group.

We consider three common anomaly detection methods: K-means [20], PCA [1] and Spectre [17] (where Spectre is the current state-of-the-art) and we report the detection results from the three defenses. Appendix A.12 gives details of these methods. The detection results on the zero-activation attack are given in Figure 11 in the appendix. Anomaly detection is very effective at identifying the samples with target property. For example, for the gender recognition and smile detection tasks, the detection rate is over 80% in most cases. These results motivate the design of stealthier attacks which we describe next.

8.2. Stealthier Model Manipulation

To evade the defense that checks the distribution of activations, we modify our zero-activation attack to ensure: (1) secreting activations for samples without the property are also non-zero (bypassing simple defense of checking abnormal zeros); (2) secreting activations of samples with and without target property are still distinct (the attack is still effective); (3) that distinction between activations should not be captured by anomaly detection methods (evading anomaly detection); (4) the actual distribution of activations that matches the attacker's goal cannot be easily guessed by the defender (handling cases when the defender actively searches other patterns in the distribution of activations).

For (1) and (2), we adapt the loss in Equation 3 as

$$\begin{cases} \alpha \cdot \max(\|f(\boldsymbol{x}) \circ \boldsymbol{m}\| - \|f(\boldsymbol{x}) \circ \neg \boldsymbol{m}\|, 0) & \text{if } y_t = 0 \\ \beta \cdot \max(\lambda \cdot \|f(\boldsymbol{x}) \circ \neg \boldsymbol{m}\| - \|f(\boldsymbol{x}) \circ \boldsymbol{m}\|, 0) & \text{if } y_t = 1 \end{cases} \quad (4)$$

where $\lambda \geq 1$. (1): The case of $y_t = 0$ is redefined to bypass the detection of abnormal zeros. Minimizing this new loss ensures that samples without the target property will have secreting activations ($f(\boldsymbol{x}) \circ \boldsymbol{m}$) with (close-to-normal) non-zero values. (2): to ensure the property is still detectable, we actively increase the difference between the secreting activations of samples with and without property. We observe that, for upstream models with reasonable performance on the main task, non-secreting activations ($f(\boldsymbol{x}) \circ \neg \boldsymbol{m}$) have similar amplitude regardless of the fed samples containing target property. Therefore, for samples with target property, as long as we ensure the secreting activations have a larger amplitude than that of non-secreting activations, there will be a distinction between secreting activations of samples with and without property. We do this by assigning larger values to λ (e.g., $\lambda \geq 1$, instead of the original $\lambda > 0$) for the second line of Equation 4 to induce sharper distinction between samples with and without property and enable higher inference performance.

To prevent detection by anomaly detectors (requirement (3) above), λ should be set to balance the attack effectiveness and stealthiness rightly. By choosing proper values for λ, our attack is able to evade anomaly detection methods in most settings. However, in some settings (mostly in gender recognition tasks), state-of-the-art anomaly detection (Spectre) can still identify most of the samples with target property. To counter this, we add an additional regularization term (weighted by parameter γ) to the overall loss function $l(\boldsymbol{x}, y, y_t)$ in Equation 2 that further improves attack stealthiness while still maintaining relatively high at-

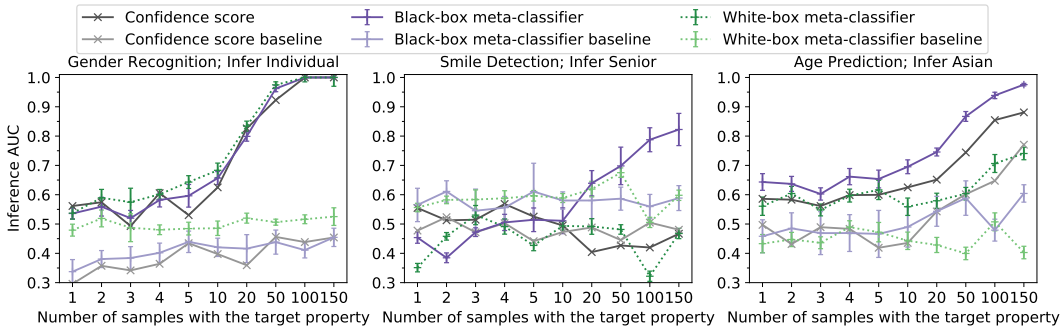

Figure 2. Inference AUC scores of the stealthier design. Since the secreting activations are no longer zero, the inference methods based on difference or variance tests are no longer applicable. The inference results of specific individuals for smile detection and age prediction also show similar improvement compared to the baseline settings (Figure 19 in the appendix). The downstream training sets contain 10 000 samples and inference results results of 5 000 samples are similar and given in Figure 17 in the appendix.

tack effectiveness. Specifically, we first obtain the corresponding covariance matrices of the activations of samples with the target property ($\boldsymbol{cov_w}$), activations of all samples with and without the target property ($\boldsymbol{cov_{w,wo}}$), and activations of samples without the target property ($\boldsymbol{cov_{wo}}$) respectively. Then, we encourage $mean(\boldsymbol{cov_w}) = mean(\boldsymbol{cov_{w,wo}}) = mean(\boldsymbol{cov_{wo}})$ and $var(\boldsymbol{cov_w}) = var(\boldsymbol{cov_{w,wo}}) = var(\boldsymbol{cov_{wo}})$ (both $mean(\cdot)$ and $var(\cdot)$ treat the whole covariance matrix as a flattened array and return scalar values) for the three covariance matrices by minimizing their differences in their mean and variance. Using this method, we ensure the distributions of activations of samples with target property will be similar to the ones without the property, making the manipulations harder to detect. We use this approach for all the experiments. To ensure the distributional pattern related to the attacker goal cannot be easily guessed (requirement (4)), we generate $\boldsymbol{m}$ randomly (instead of picking first $\|\boldsymbol{m}\|$ activations in Section 7). This makes the brute-force search of possible patterns computationally infeasible (details in Appendix A.14).

8.3. Experiments with Stealthy Attacks

Detection Evasion. Figure 16 (in the appendix) summarizes the results of our experiments to detect the stealthy upstream models (Appendix A.13 provides details on these experiments). We find that the anomaly detection methods are ineffective against our stealthier attack— $< 10\%$ of samples with the target property are detected across all settings with the exception of a detection rate $< 20\%$ (still low) for smile detection when the total number of samples is 5 000 and 100 or 150 of them are with the target property. We also made several attempts to approximately identify (instead of brute-force search) possible attack patterns in the activations but none of these succeeded in uncovering the stealthy attacks (details are in Appendix A.14).

Inference Results. From Figure 2, we can see that activation manipulation still leads to significantly improved inference results compared to the baselines with normally trained upstream models. For example, for gender recognition, when ≥ 50 downstream training samples have the target property, inference AUC scores exceed 0.95, which is a huge improvement compared to the baseline attack where all AUC scores are less than 0.6, and similar trends follow for smile detection (with over 100 samples with property, AUC improves from < 0.6 to > 0.78) and age prediction (with over 100 samples with property, AUC improves from < 0.77 to > 0.9). Comparing the results for the stealthier attacks to the results that do not consider defenses in Figure 1, we observe that the attack effectiveness declines as expected since we are now trading-off attack effectiveness for stealthiness. Training models with the attack goal poses negligible impact on the model performance (accuracy drop $< 0.9\%$, see Appendix A.6).

9. Conclusion

Our work demonstrates how a malicious upstream trainer can manipulate its training process to amplify property inference risks for downstream models when transfer learning is done. Our empirical results show that such manipulations can be exploited to enable very precise property inference, even in black-box settings, across a variety of tasks. Although there is potential for a new arms race between methods of hiding manipulations and methods of detecting them, the larger lesson from this work, and other works exposing similar risks, is that it is important that users of pretrained models to only use models from trusted providers.

Acknowledgements

This work was supported in part by the National Key R&D Program of China (#2022YFF0604503 and #2021YFB3100300), the United States National Science Foundation through the Center for Trustworthy Machine Learning (#1804603), NSFC (#62272224), JiangSu Province Science Foundation for Youths (#BK20220772), and Lockheed Martin Corporation.

References

[1] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010. 7, 19

[2] MAH Akhand, Ijaj Sayim, Shuvendu Roy, N Siddique, et al. Human Age Prediction from Facial Image Using Transfer Learning in Deep Convolutional Neural Networks. In *International Joint Conference on Computational Intelligence*, 2020. 5

[3] Giuseppe Ateniese, Luigi V Mancini, Angelo Spognardi, Antonio Villani, Domenico Vitali, and Giovanni Felici. Hacking Smart Machines with Smarter Ones: How to Extract Meaningful Data from Machine Learning Classifiers. *International Journal of Security and Networks*, 10(3):137–150, 2015. 1, 2

[4] Franziska Boenisch, Adam Dziedzic, Roei Schuster, Ali Shahin Shamsabadi, Ilia Shumailov, and Nicolas Papernot. When the curious abandon honesty: Federated learning is not private. *arXiv preprint arXiv:2112.02918*, 2021. 2

[5] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *IEEE International Conference on Automatic Face & Gesture Recognition*, 2018. 5

[6] Shuvam Chakraborty, Burak Uzkent, Kumar Ayush, Kumar Tanmay, Evan Sheehan, and Stefano Ermon. Efficient Conditional Pre-training for Transfer Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1

[7] Melissa Chase, Esha Ghosh, and Saeed Mahloujifar. Property Inference from Poisoning. In *IEEE Symposium on Security and Privacy*, 2022. 1, 2

[8] Harsh Chaudhari, Jackson Abascal, Alina Oprea, Matthew Jagielski, Florian Tramèr, and Jonathan Ullman. SNAP: Efficient Extraction of Private Properties with Poisoning. *arXiv preprint arXiv:2208.12348*, 2022. 2

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2009. 2, 5, 12

[10] Fadi Dornaika, Ignacio Arganda-Carreras, and C Belver. Age estimation in facial images through transfer learning. *Machine Vision and Applications*, 30(1):177–187, 2019. 5

[11] Liam Fowl, Jonas Geiping, Wojtek Czaja, Micah Goldblum, and Tom Goldstein. Robbing the fed: Directly obtaining private data in federated learning with modified models. *arXiv preprint arXiv:2110.13057*, 2021. 2

[12] Karan Ganju, Qi Wang, Wei Yang, Carl A Gunter, and Nikita Borisov. Property Inference Attacks on Fully Connected Neural Networks using Permutation Invariant Representations. In *ACM SIGSAC Conference on Computer and Communications Security*, 2018. 1, 2, 5, 6

[13] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics*, 2010. 5

[14] Kathrin Grosse, Thomas A Trost, Marius Mosbach, and Michael Backes. Adversarial initialization-when your network performs the way I want. *arXiv preprint arxiv:1902.03020*, 2019. 2

[15] Xin Guo, Luisa Polania, and Kenneth Barner. Smile Detection in the Wild Based on Transfer Learning. In *IEEE International Conference on Automatic Face & Gesture Recognition*, 2018. 5

[16] Valentin Hartmann, L'eo Meynent, Maxime Peyrard, Dimitrios Dimitriadis, Shruti Tople, and Robert West. Distribution inference risks: Identifying and mitigating sources of leakage. *arXiv preprint arXiv:2209.08541*, 2022. 2

[17] Jonathan Hayase, Weihao Kong, Raghav Somani, and Sewoong Oh. SPECTRE: Defending Against Backdoor Attacks Using Robust Statistics. In *International Conference on Machine Learning*, 2021. 7, 19, 20, 23

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *IEEE International Conference on Computer Vision*, 2015. 5

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 3, 5

[20] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data Clustering: A Review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999. 7, 19

[21] Joanna Jaworek-Korjakowska, Pawel Kleczek, and Marek Gorgon. Melanoma Thickness Prediction Based on Convolutional Neural Network with VGG-19 Model Transfer Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 1

[22] Marc Juárez, Samuel Yeom, and Matt Fredrikson. Black-Box Audits for Group Distribution Shifts. *arXiv preprint arXiv:2209.03620*, 2022. 1, 2

[23] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting Unintended Feature Leakage in Collaborative Learning. In *IEEE Symposium on Security and Privacy*, 2019. 2

[24] Cao Hong Nga, Khai-Thinh Nguyen, Nghi C Tran, and Jia-Ching Wang. Transfer Learning for Gender and Age Prediction. In *IEEE International Conference on Consumer Electronics-Taiwan (ICCE-Taiwan)*, 2020. 5

[25] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 5

[26] Roei Schuster, Tal Schuster, Yoav Meri, and Vitaly Shmatikov. Humpty Dumpty: Controlling Word Meanings via Corpus Poisoning. In *IEEE Symposium on Security and Privacy*, 2020. 2

[27] Ozan Sener and Vladlen Koltun. Multi-Task Learning as Multi-Objective Optimization. In *Advances in Neural Information Processing Systems*, 2018. 13

[28] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership Inference Attacks Against Machine Learning Models. In *IEEE Symposium on Security and Privacy*, 2017. 4

[29] Anshuman Suri and David Evans. Formalizing and Estimating Distribution Inference Risks. In *Privacy Enhancing Technologies Symposium*, 2022. 1, 2, 4, 5, 6

[30] Anshuman Suri, Pallika Kanani, Virendra J Marathe, and Daniel W Peterson. Subject Membership Inference Attacks in Federated Learning. *arXiv preprint arXiv:2206.03317*, 2022. 2

[31] Philipp Terhörst, Daniel Fährmann, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. MAAD-Face: A Massively Annotated Attribute Dataset for Face Images. *IEEE Transactions on Information Forensics and Security*, 16:3942–3957, 2021. 5, 11, 12

[32] Philipp Terhörst, Marco Huber, Jan Niklas Kolf, Ines Zelch, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Reliable Age and Gender Estimation from Face Images: Stating the Confidence of Model Predictions. In *IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2019. 5, 11

[33] Bolun Wang, Yuanshun Yao, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. With Great Training Comes Great Vulnerability: Practical Attacks against Transfer Learning. In *USENIX Security Symposium*, 2018. 1, 2, 3, 5

[34] Xiuling Wang and Wendy Hui Wang. Group Property Inference Attacks Against Graph Neural Networks. *arXiv preprint arXiv:2209.01100*, 2022. 2

[35] Yuxin Wen, Jonas A Geiping, Liam Fowl, Micah Goldblum, and Tom Goldstein. Fishing for user data in large-batch federated learning via gradient magnification. In *International Conference on Machine Learning*, 2022. 2

[36] Yu Xia, Di Huang, and Yunhong Wang. Detecting Smiles of Young Children via Deep Transfer Learning. In *IEEE International Conference on Computer Vision Workshops*, 2017. 5

[37] Xiaojun Xu, Qi Wang, Huichen Li, Nikita Borisov, Carl A Gunter, and Bo Li. Detecting AI Trojans Using Meta Neural Analysis. In *IEEE Symposium on Security and Privacy*, 2021. 2, 4

[38] Kaiyu Yang, Jacqueline Yau, Li Fei-Fei, Jia Deng, and Olga Russakovsky. A Study of Face Obfuscation in ImageNet. In *International Conference on Machine Learning*, 2022. 12

[39] Yuanshun Yao, Huiying Li, Haitao Zheng, and Ben Y Zhao. Latent Backdoor Attacks on Deep Neural Networks. In *ACM SIGSAC Conference on Computer and Communications Security*, 2019. 1, 2, 3, 5

[40] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*, 2018. 11

[41] Wanrong Zhang, Shruti Tople, and Olga Ohrimenko. Leakage of Dataset Properties in Multi-Party Machine Learning. In *USENIX Security Symposium*, 2021. 1, 2, 4

[42] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A Comprehensive Survey on Transfer Learning. *Proceedings of the IEEE*, 109, 2020. 1

[43] Yang Zou, Zhikun Zhang, Michael Backes, and Yang Zhang. Privacy Analysis of Deep Learning in the Wild: Membership Inference Attacks against Transfer Learning. *arXiv preprint arXiv:2009.04872*, 2020. 2

A. Appendix

A.1. Overcoming Training Data Limitations

Due to the possible inadequacies of representative samples in the upstream training data, practical implementation with good performance can be challenging. Below, we discuss the three main challenges in crafting the pretrained model in practice, and our ways of addressing them.

Imbalance between Samples with and without Target Property. If the upstream training set contains a large number of samples with only a small fraction with the target property, optimization of the loss function related to samples with the target property (Second line of Equation 3) can have convergence issues. To deal with this scenario, we use mixup-based data augmentation to increase the number of samples with the target property in the upstream training set [40]. Additionally, to reduce the training time (faster convergence) for the upstream model, we also use a clean pre-trained model as the starting point for obtaining the final manipulated model.

Lack of Upstream Labels for Samples with Target Property. If samples with the target property are already present in the upstream training set, the attacker can directly train its model using Equation 2. However, this may not always be the case in practice and the attacker may need to inject additional samples with the target property (that are available to the attacker), with the label information for these injected samples being unavailable. For example, if the target property is a specific individual, when adding the images of that individual to ImageNet dataset, we may not be able to find proper labels for injected images out of the original 1K possible labels. However, these labels are required for optimizing l_{normal}. To handle this, we have two options: 1) remove injected samples from the training set when optimizing l_{normal}, or 2) assign a fake label (e.g., create a fake $n+1$ label for injected samples in a n-class classification problem) and remove parameters related to the fake label in the final classification layer before releasing models. The first option has negligible impact on the main task accuracy in all settings, but resultant attack effectiveness is inferior to the second one. In contrast, the second option usually gives better inference results, but in some settings (e.g., experiments when pretrained models are face recognition models in Section 7), can have non-negligible impact on the main task accuracy. Therefore, we choose the second option when it does not impact the main task performance much and switch to the first one when it does.

Lack of Representative Non-Target Samples in Training Set. The space of samples without the target property can be much larger than the space of samples with the target property as the former can contain combinations of multiple data distributions. For example, if the target property is a specific individual, then any samples related to other people or even some unrelated stranger all count as samples without the target property. However, in practice, the upstream trainer's data may not contain enough non-target samples to be representative. This can be a problem when minimizing the loss item related to the samples without the target property (first line of Equation 3), as secreting activations may not be sufficiently suppressed for those samples. To solve this, we choose to augment upstream training set with some representative samples without the target property and name this method as *Distribution Augmentation*. For example, when the target property is a specific person, the attacker can inject samples of new people not present in the current upstream training set and thus expand the upstream distribution. The labels for these newly injected samples are handled similarly to the labels for additionally injected samples with target property. An ablation study on the importance of distribution augmentation is given in Appendix A.9.

A.2. Details of Dataset Settings

As introduced in Section 6, we experiment with three transfer learning tasks: gender recognition, smile detection, and age prediction. We consider the property inference of determining whether images of specific individuals are present in the downstream training set for all these tasks. And for the smile detection and age prediction, we consider additional inference targets: inferring the presence of senior people for smile detection and the presence of Asian people for age prediction. As for the inference of the existence of specific individuals, we choose the person who has the most samples in VGGFace2 as the inference target for both gender recognition and age prediction, and choose the person who has the most samples of smile labels (provided by MAADFace [31, 32]) as the target for smile detection (the person with the most samples in VGGFace2 does not have enough samples with valid labels for the smile attribute). We choose the target property in this manner mainly for convenience in conducting experiments, as the upstream model training, victim model training, and shadow model training (for meta-classifier-based property inference) (ideally) require no overlaps between their training data to mimic the hardest attack scenario. Subsequently, if we choose a target with small number of samples in the original dataset, then we may have trouble in performing the three types of model training effectively.

In the upstream training, since we use the techniques described in Appendix A.1, we need to inject samples with and without the target property into the original upstream training set. And for the downstream model training, we first prepare downstream candidate sets based on VGGFace2 and then construct various downstream settings using the samples from the candidate sets (Appendix A.3).

Task	Target Property	Samples injected into Upstream training		Downstream Candidate set	
		w/ property	w/o property	w/ property	w/o property
Gender Recognition	Specific Individuals	342	1 710	250	200 000
Smile Detection		261	1 305	250	200 000
Age Prediction		342	1 710	250	165 915
Smile Detection	Senior	3 000	15 000	1 000	200 000
Age Prediction	Asian	3 000	15 000	1 000	128 528

Table 2. Number of samples injected into the upstream training and in the downstream candidate sets

Table 2 summarizes the number of samples of the sample injection and the downstream candidate sets. The details of the three transfer learning tasks are reported below:

Gender recognition. We randomly select 50 people from VGGFace2 and train face recognition models classifying those 50 people as the upstream model. For each person, we randomly choose 400 samples for training and 100 for testing. To avoid overlap, we also ensure that any images of these 50 people do not appear in the downstream training. Since the individual targeted by the adversary (the inference target) is not in the randomly chosen upstream set, we inject 342 randomly chosen samples with the target property into the upstream training set to achieve the attack. Note that, we also need to assign enough disjoint samples with the target property to the downstream training and meta-classifier training, and 342 is the maximum number of samples that we can assign to the upstream training as there are limited samples with the target property in VGGFace2. For the distribution augmentation described in Appendix A.1, we inject 1 710 samples (5×342) without the target property to the upstream set, and those injected samples are randomly sampled from VGGFace2 and are from individuals that are not in the original upstream training set. As for the downstream candidate set, there are 250 samples with the target property and 200 000 samples without the target property. All the samples in the candidate set are randomly sampled from VGGFace2 and have no overlap with those in the upstream training.

Smile detection. We have two inference targets for this transfer learning task. For the inference of the specific individual, the number of samples with the target property injected into the upstream set is 261 (number decreased compared to gender recognition since there are fewer samples with the target property in VGGFace2 for this inference task), and the number of samples without the target property for distribution augmentation is 1 305 (5×261). The candidate set for the downstream training has 250 samples with the target property and 200 000 samples without the target property.

As for the inference of the presence of senior people, since there are plenty of samples labeled as seniors in VGGFace2 [31], we increase the number of samples injected into the upstream training set and inject 3 000 samples with the target property and 15 000 samples without the target property (distribution augmentation). The original upstream training set is ImageNet [9]. However, ImageNet contains images of human beings, and there are no "senior" labels for those images. Instead of manually labeling them, we remove all the facial images in ImageNet for this inference task. We use the facial labels provided by Yang et al. [38] when conducting the removing. The downstream candidate set has 1 000 samples (number increased since there are more samples available) with the target property and 200 000 samples without the target property.

Age prediction. We also have two inference targets for this transfer learning task. For the inference of the presence of the specific individual, the numbers of samples with and without the target property injected into the upstream training set are 342 and 1 710 respectively, which are the same as those in the gender recognition task as the target properties are the same in these two tasks. The downstream candidate set has 250 samples with the target property and 165 915 samples without the target property.

As for the inference of the presence of Asian people, we inject 3 000 samples with the target property (Asian) and 15 000 samples without the target property into the upstream training set. These two numbers are the same as those in the smile detection task with senior people as the target property. We also remove all the facial images in ImageNet for this inference task. The downstream candidate set has 1 000 samples with the target property and 128 528 samples without the target property. The number of samples without the target property in the downstream candidate set in the age prediction task is less than those in other settings. This is because we are not able to find enough samples with valid ethnic labels using the attribute labels provided by MAADFace.

A.3. Details of Downstream Training and Adversary's Meta-Classifier Training

As described in Appendix A.2, to generate the downstream training set, we first prepare randomly selected sam-

ples without the target property and samples with the target property to form the downstream candidate set, and then construct downstream sets based on the candidate set. Specifically, a downstream training set of size n is generated by randomly sampling from this candidate set while also specifying the number of samples with target property as n_t. For experiments in this section, we consider settings where $n = 5\,000$ or $10\,000$, and n_t takes value from $\{0, 1, 2, 3, 4, 5, 10, 20, 50, 100, 150\}$ (this gives $2 \times 11 = 22$ different settings). We train 32 downstream models with different random seeds for each setting, and those models will be used for computing inference AUC scores (the models trained with $n_t = 0$ are used as the reference group).

To train the meta-classifier attacks, the attacker needs to train many downstream shadow models and thus, we also prepare a separate downstream candidate set with the same size as the victim's downstream candidate set but without any overlaps on the data. This simulates the most difficult and realistic scenario for the attacker. We also ensure that no samples in the two downstream candidate sets appear in the upstream training set, which again makes the attack more difficult. To simulate the victim's downstream training, we assume the attacker also uses a downstream training set of size n, but has no overlap with the actual victim's downstream training set. In Appendix A.8, we relax this assumption and show our attack retains its effectiveness even when the size of the victim's downstream training dataset is unknown to the adversary. For each setting with fixed n, the attacker trains 320 shadow downstream models (256 for training, 64 for validation) for each of the distributions (with and without target property). The number of training samples with the target property for each model is randomly selected from the range $[1, 170]$, which simulates the scenario where the value of n_t of the victim downstream model cannot be accurately guessed.

A.4. Baseline Results

In this section, we focus on experiments where the upstream model is trained normally, without considering the attack goals described in Section 4 and Section 8.2. For these baseline experiments, there are no secreting parameters (i.e., manipulated secreting activations) in the model, so the attacker can only use the attacks that are not directly related to the manipulation.

We experiment with the confidence score test, the black-box meta-classifier, and the white-box meta-classifier, and report AUC scores for distinguishing between models trained with and without the target property. For meta-classifier-related inferences, we report the average AUC values over five runs of meta-classifiers with different random seeds, along with their standard deviation. Figure 3 shows the results. We observe that the attacks have inference AUC scores less than 0.82, with most (4 out of 6 settings) of them with scores less than 0.7. Moreover, we do not find a clear winner from the three inference methods we test. These results demonstrate the limited effectiveness of existing methods applicable to normally trained upstream models.

A.5. Hyperparameter Setup of Zero-Activation Attacks

In Section 7, when training upstream models for the zero-activation attack (Section 4), we set α and β to 1, treating all loss terms equally. We tried different settings on α and β, as well as methods that automatically set them [27], but no significant improvements are observed, so we just use those simplest choices. We also tested different values for λ and $\boldsymbol{m}$, but did not observe significant differences in the attack effectiveness, suggesting our attack is not sensitive to hyperparameters. Details of experiments on different combinations of λ and $\boldsymbol{m}$ are in Appendix A.7. For the results in Section 7, we select λ values that are big enough while ensuring the upstream model accuracy is not impacted significantly ($\lambda = 10$ for smile detection and age prediction, and $\lambda = 5$ for gender recognition). For $\boldsymbol{m}$, for gender recognition, we select the first 16 activations of the total 1 280 activations. For smile detection and age prediction, since the first layer of downstream model is convolutional, we can only select activations at the granularity of channels, and we choose to manipulate the first channel of the total 256 channels. We also use the distribution augmentation described in Appendix A.1 in the upstream training; ablation studies (Appendix A.9) suggest it is crucial for performance.

A.6. Impact of Activation Manipulation to Model Accuracy

Upstream model accuracy. We find that the upstream training accuracy will not be significantly affected by the manipulation. Table 3 shows the accuracy drop is less than 0.9% for the attacks used in Section 7 and Section 8.3. For different hyperparameter settings of the zero-activation attack, Table 4 shows that the accuracy of the upstream models will drop by at most 1.9% for all the settings except the upstream models of the gender recognition task when λ is too high (10 or 20). The possible explanation is that the MobileNetV2 architecture used in those settings does not have enough capacity for achieving the difference (between activations of the samples with and without the target property) defined by λ while maintaining high task accuracy.

Downstream model accuracy. The downstream model accuracy is not affected by the attack either. Table 3 shows the averaged accuracy of the downstream models (excluding the downstream models trained for preparing meta-classifiers) trained in Section 7 and Section 8.3. We do not observe any accuracy drop brought by the attack, instead

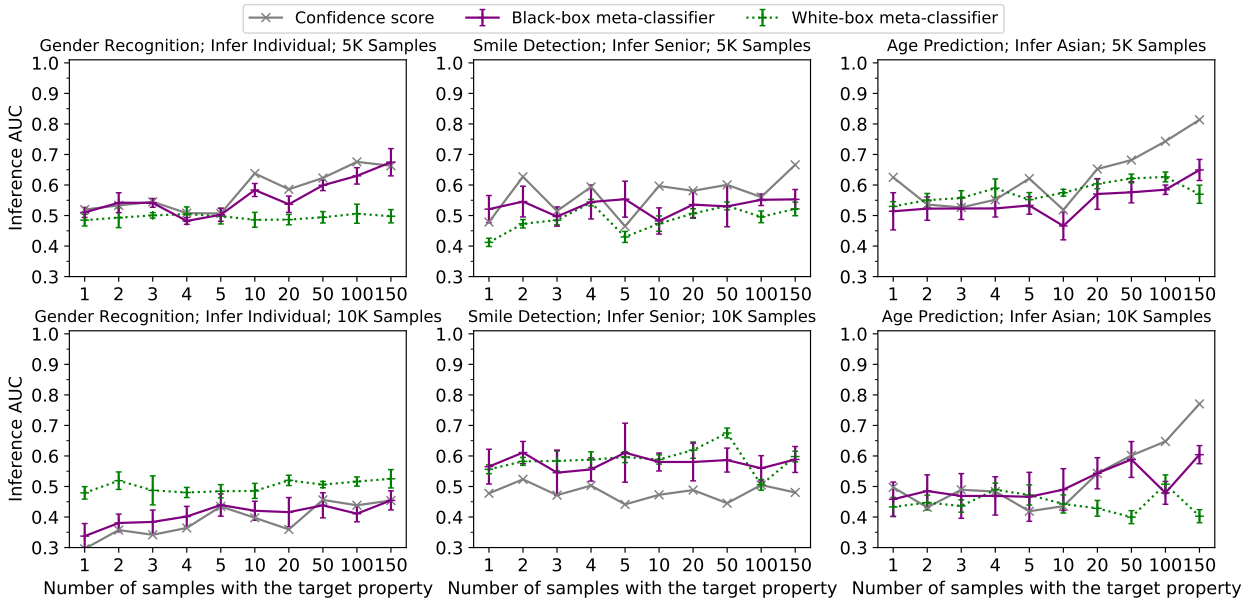

Figure 3. Inference AUC scores when upstream models are trained normally. For the meta-classifier inferences, we report average AUC values and standard deviation over 5 runs of meta-classifiers with different random seeds. For normally trained models, only the inference attacks that are not directly related to the manipulation are applicable. The first and second rows show results when downstream training sets contain 5 000 and 10 000 samples respectively. Results of the inference of specific individuals for smile detection and age prediction show similar trends and are found in Figure 14.

Task	Target Property	Upstream Accuracy			Downstream Accuracy		
		Clean Model	Zero-Activation Attack	Stealthier Attack	Clean Model	Zero-Activation Attack	Stealthier Attack
Gender Recognition	Specific Individuals	92.8	92.6	92.1	95.7 (95.8)	95.8 (95.8)	95.7 (95.8)
Smile Detection		73.2	73.5	73.5	90.0 (90.5)	90.4 (90.8)	90.2 (90.7)
Age Prediction		69.7	70.1	70.2	91.4 (92.4)	91.6 (92.5)	91.6 (92.6)
Smile Detection	Senior	73.2	72.5	72.7	88.3 (88.9)	88.8 (89.4)	88.8 (89.3)
Age Prediction	Asian	69.7	68.8	69.1	91.4 (92.5)	91.5 (92.6)	91.6 (92.7)

Table 3. Upstream and downstream model accuracy. The clean models are the models trained without attack goals (manipulation), and for smile detection and age prediction, we directly use the pretrained ImageNet models released by PyTorch as the clean upstream models. For the downstream accuracy, we report the averaged accuracy of the downstream models (excluding the downstream models trained for preparing meta-classifiers) trained in Section 7 and Section 8.3. The values outside the parenthesis are the averaged accuracy for the downstream models that are trained with 5 000 samples, while the values inside the parenthesis are the results for the 10 000 samples.

all the accuracies are slightly improved after manipulation. Currently, we are unclear about the root cause for this observation and will leave the detailed exploration on this as future work.

A.7. Impact of Hyperparameters

This section explores the impact of the hyperparameters, λ and $\boldsymbol{m}$, in the loss function of upstream model training in Equation 3, to the effectiveness of the zero-activation attack.

Impact of λ. The hyperparameter λ in Equation 3 is directly related to the magnitude of the difference between the downstream models trained with and without the target property and therefore, is critical to the effectiveness of the inference attacks (larger λ generally means more effective attacks). In this section, we compare the inference effectiveness on downstream models when the upstream models are trained with different λ values. Since training the upstream models are costly, we only choose λ from $\{1, 5, 10, 20\}$. For the inference method, for each task, we select the best performing white-box inference attacks—for the gender recognition task, we choose the variance test (parameter difference test is not available for this task) and for the other two tasks, we choose the parameter difference test, and report

Task	Clean Model	Zero-Activation Atatck							
		λ				$\|\|\boldsymbol{m}\|\|_1$			
		1	5	10	20	8/1C	16/4C	32/8C	64/16C
Gender Recognition (Infer Individual)	92.8	92.5	92.6	90.3	64.1	93.2	92.6	92.5	92.8
Smile Detection (Infer Senior)	73.2	72.7	72.7	72.5	72.1	72.5	72.6	72.7	72.5
Age Prediction (Infer Asian)	69.7	69.1	69.0	68.8	67.8	68.8	68.8	68.7	68.7

Table 4. Upstream model accuracy of zero-activation attacks for different hyperparameter settings. We vary the values of λ or $\|\boldsymbol{m}\|_1$ in the experiments and use the remaining experimental settings in Appendix A.5.

the results in Figure 4. We also conducted experiments using black-box inference methods and results are included in Figure 5. The rest of the settings are the same as those used in Section 7.

Figure 4 gives the white-box inference results. For the gender recognition and age prediction tasks, by comparing different lines corresponding to different λ values, the general trend is if we increase λ, the inference AUC scores will first (expectedly) increase and then decrease. For example, for gender recognition, increasing λ from 1 to 5, the AUC scores are consistently improved in all settings with varying number of target samples in the downstream training set (the average AUC score increases from 0.84 to 0.94). But further increasing λ to 10 and 20 does not help and the inference performs consistently worse as λ gets larger (e.g., average AUC score drops from 0.89 of $\lambda = 5$ to 0.50 of $\lambda = 20$). In contrast, for smile detection task, the inference performance continues to increase as we increase λ in general. For all the tasks, we initially observe increased attack effectiveness by increasing λ because larger λ makes the distinction between downstream models trained with and without property more significant and hence is easier for the subsequent inference attacks. But when λ gets too large, for settings where the inference effectiveness decreases, we observe that the loss function related to the attacker goal ($l_t(\cdot)$ in Equation 2) starts to interfere with the main task training ($l_{normal}(\cdot)$) and fails to converge at the end of upstream training (Table 4). For smile detection, $l_t(\cdot)$ still converges well (may be because the upstream model has enough capacity) and hence the inference effectiveness continues to increase as the increase of λ.

In Figure 4, although the choice of λ does have some impact on the inference effectiveness, we find that our attack still works quite well for a wide range of λ values. For example, for gender recognition, AUC scores are quite high and exceed 0.9 if $\geq$ 10 samples are with the target property when the value of λ is between 1 and 10; for the other two tasks, when the value of λ is between 5 and 20, AUC scores also exceed 0.9 if $\geq$ 20 samples are with the target property. We have similar observations as above (i.e., the trend of inference effectiveness as λ changes and good attack performance for a wide range of λ) when we replace the white-box inference methods with black-box ones and details can be found in Figure 5.

Impact of $\boldsymbol{m}$. The hyperparameter $\boldsymbol{m}$ controls the location and number of activations selected for manipulation in Equation 3. We empirically find that, with the same size of activations $\|\boldsymbol{m}\|_1$, the location of $\boldsymbol{m}$ does not have a significant impact on attack effectiveness, and therefore, we fix the selection of manipulated activations to be the first n_t activations (i.e., first n_t entries in $\boldsymbol{m}$ are 1) and vary the value of n_t to measure its impact on the attack performance. The rest of the experimental settings are the same as in Section 7. We choose the first 8, 16, 32 and 64 of the total 1 280 activations as the secreting activations for the gender recognition task. For the smile detection and the age prediction tasks, we select the first 1, 4, 8, and 16 channels out of 256 channels as the secreting activations.

The inference methods adopted are the same as those in the study of the impact of λ and the white-box results are reported in Figure 6. From the figure, we observe that, in general, the inference effectiveness increases as we increase the number of selected activations (i.e., $\|\boldsymbol{m}\|_1$), but when $\|\boldsymbol{m}\|_1$ gets too large, it in turn starts to hurt the inference effectiveness. The possible reason is still similar to the one in the study of the impact of λ: initially, when more activations are selected for manipulation, the difference between the downstream models trained with and without the target property will be more significant, and makes the subsequent inference attacks more effective. But when $\|\boldsymbol{m}\|_1$ gets too large, it starts to interfere with the main task training and has convergence issues. From Figure 6, we also observe that the inference AUC scores remain high across all selections of $\boldsymbol{m}$. For example, AUC scores are all > 0.9 when $\geq$ 20 downstream training samples have the target property for gender recognition and smile detection and when $\geq$ 50 downstream training samples are with the target property for age prediction. Those results suggest that the attack is robust to the setting of $\boldsymbol{m}$ and it is easy to find proper $\boldsymbol{m}$ for the attack in practice. Similar observations are also found when we replace the white-box inference methods with black-box ones (details in Figure 7).

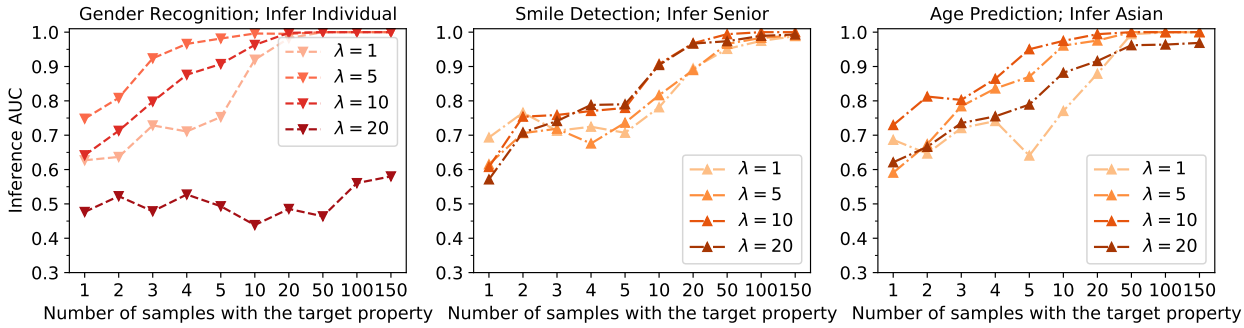

Figure 4. Inference AUC scores of white-box methods for different values of λ (Equation 3). All downstream training sets have 5 000 samples. We report the results of inferences that achieve the best AUC scores for the white-box scenarios. Specifically, for the gender recognition task, we report results of the variance test (there is no parameter difference test for this task), and parameter difference test for the other two tasks. Results of the black-box inferences show a similar trend (Figure 5).

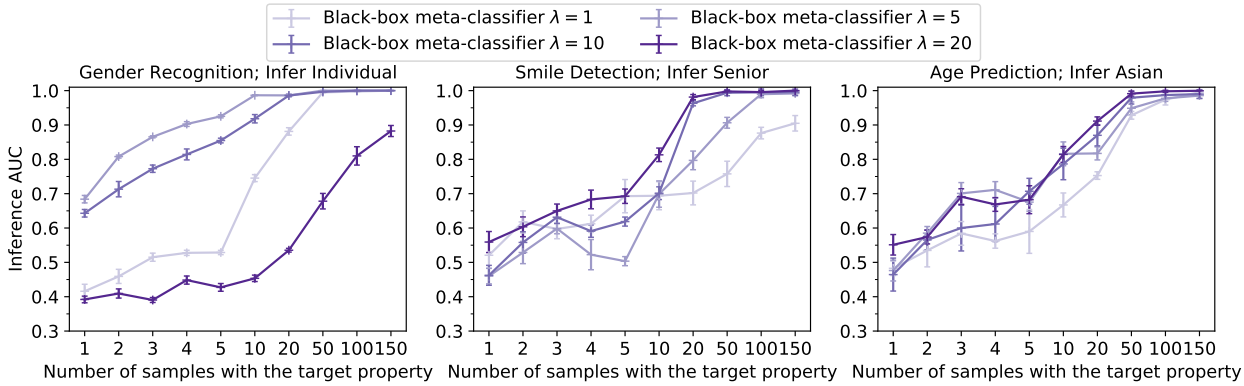

Figure 5. Inference AUC scores of black-box inferences for different values of λ (Equation 3). All the downstream training sets have 5 000 samples in these results. We only report the results of the better performing black-box inference method (i.e., the black-box meta-classifiers) here. The results of the white-box attacks show a similar trend and can be found in Figure 4.

A.8. Impact of the knowledge of the size of the downstream set

In Section 7, when conducting property inference with meta-classifiers, the attacker trains shadow models using the same downstream training set size n as the victim. In this section, we show that, for meta-classifier-based attacks, the knowledge of downstream training size used by the victim does not impact inference effectiveness much.

In the experiments, we fix the size of the victim training set to 5 000 (i.e., $n = 5\,000$) and vary the sizes of the (simulated) downstream training sets of the attacker. Specifically, we set the attacker training size to 2 500, 5 000, 7 500, and 10 000 separately and remaining experimental setups are kept the same as in Section 7.

Figure 8 shows the inference results of the meta-classifier-based approaches. For both the white-box and black-box methods, varying the training set size has negligible impact on the inference performance: for the black-box approach, the purple lines stay very close to each other and the AUC scores all exceed 0.8 when $\geq$ 20 samples out of the total 5 000 samples have the target property and exceed 0.95 when $\geq$ 50 samples are with the property. Similarly, for the white-box meta-classifiers approach, the green lines also stay close to each other and the AUC scores all exceed 0.9 when $\geq$ 100 samples have the target property.

A.9. Importance of Distribution Augmentation

In Appendix A.1, we introduce distribution augmentation for upstream training, which injects representative samples without the target property into the upstream training set to better achieve the attack goal described in Equation 3. Figure 9 shows the attack performance when we do not use distribution augmentation. The victim training set size is set to 5 000 and other experimental setups are the same as those in Section 7. From the figure, we observe that AUC scores of attacks without distribution augmentation are all less than 0.86, and get even lower (< 0.7) for gender recognition and smile detection. These scores are significantly lower than the results with distribution augmentation (de-

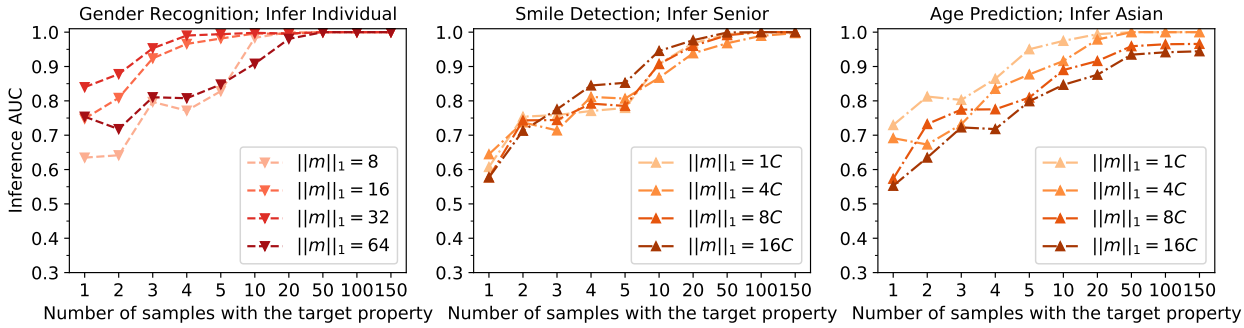

Figure 6. Inference AUC scores of of white-box methods for different number of activations (the $\boldsymbol{m}$ in Equation 3). All downstream training sets have 5 000 samples. We only report results of inferences that achieve the best AUC scores (variance test for gender recognition and parameter difference test for the other two tasks). Results of the black-box inferences show a similar trend (Figure 7).

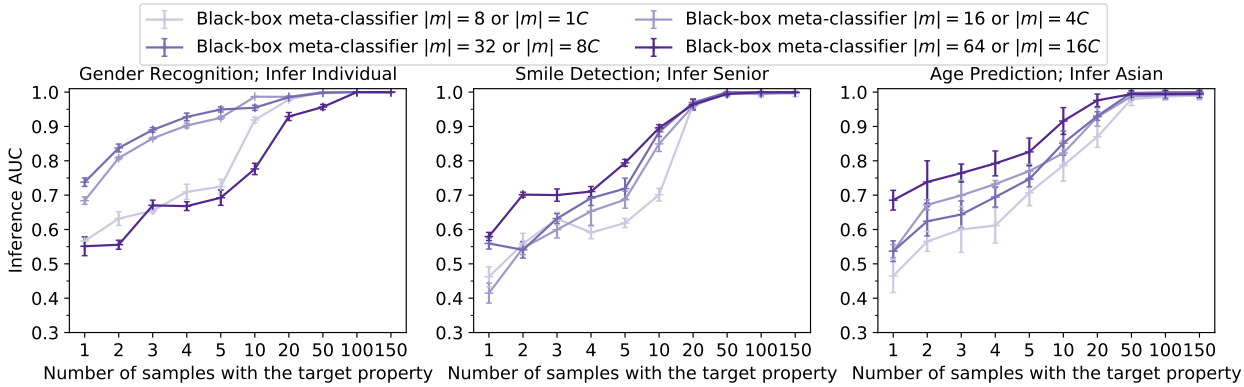

Figure 7. Inference AUC scores of black-box inferences for manipulating different number of activations (the $\boldsymbol{m}$ in Equation 3). All the downstream training sets have 5 000 samples in these results. We only report the results of the better performing black-box inference method (i.e., the black-box meta-classifiers) here. The results of the white-box attacks show a similar trend and can be found in Figure 6.

tails in Figure 12 and 1). For example, with the augmentation, AUC scores all exceed 0.9 if more than 20 samples are with the target property and the importance of distribution augmentation is thus apparent.

A.10. AUC values < 0.5

We observe that a few attack settings have AUC scores consistently below 0.5. Those rare abnormal AUC scores mainly occur for black-box methods against normal pretrained models (e.g., the confidence score test and black-box meta-classifier for the gender recognition with 10 000 downstream samples in Figure 3.) For the confidence score test, by manual inspection, we find its working assumption is not satisfied by the downstream models fine-tuned from normal pretrained models in some settings. The confidence score test assumes models trained with the property perform better on samples with the property than those trained without the property, but an opposite pattern is observed for the queried downstream models. As for black-box meta-classifiers, we observe the anomalies happen when the inference tasks are too challenging and the meta-classifiers cannot obtain meaningful information but overfit to the training set (despite early stopping). Specifically, AUC scores are high (> 0.75) on the training set, ~ 0.5 on the validation set, and show anomalies (< 0.5) on the test set. We note that the gap between the validation set and the test set is large because they are trained differently. When training downstream models with the target property for the training and validation set, we randomly sample 1-170 samples with the property each time to simulate the real-world case (discussed in Appendix A.3), while for the test set, we randomly sample fixed number of samples with the property for each AUC computation (e.g., 1, 2, ..., 150) to show the trend. We reemphasize that those anomalies mainly happen in the non-manipulation settings because of the limitation of inference methods on normal pretrained models when the inference tasks are too challenging. Our proposed manipulation (e.g., providing stronger signal) lowers the difficulty of those challenging cases and leads to better/normal results.

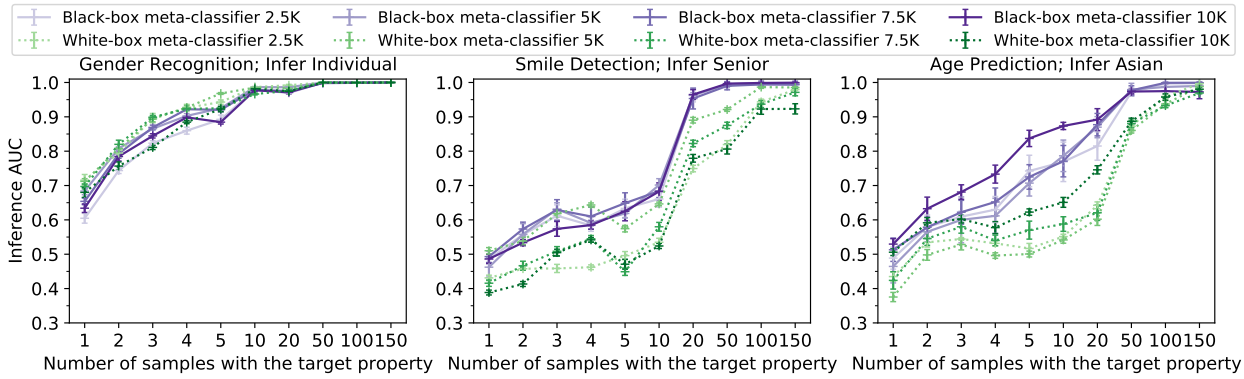

Figure 8. Inference AUC scores of meta-classifiers when the shadow models of the meta-classifiers are trained on datasets of different sizes. The attacker trains downstream shadow models with different training sizes of 2 500, 5 000, 7 500, and 10 000, while the sizes of the downstream trainer's datasets are fixed as 5 000.

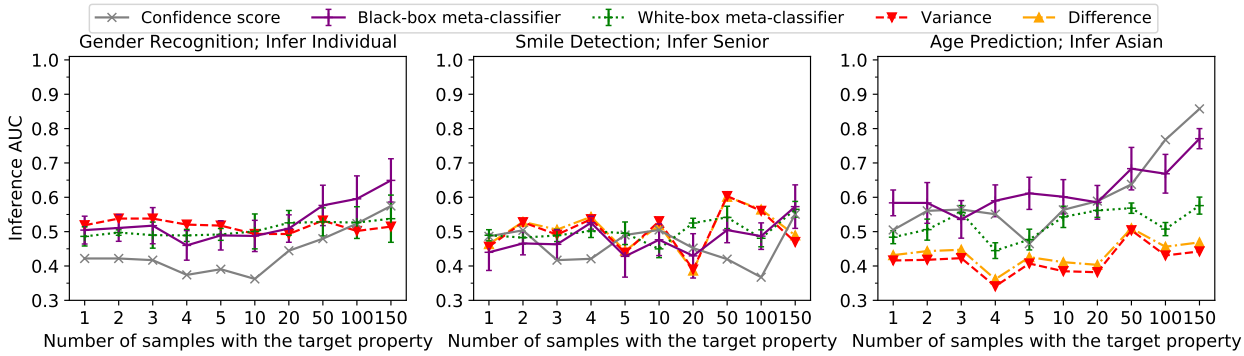

Figure 9. Inference AUC scores when upstream models are not trained with distribution augmentation (Appendix A.1). All the downstream training sets have 5 000 samples in these results.

A.11. Inferring Multiple Properties Simultaneously

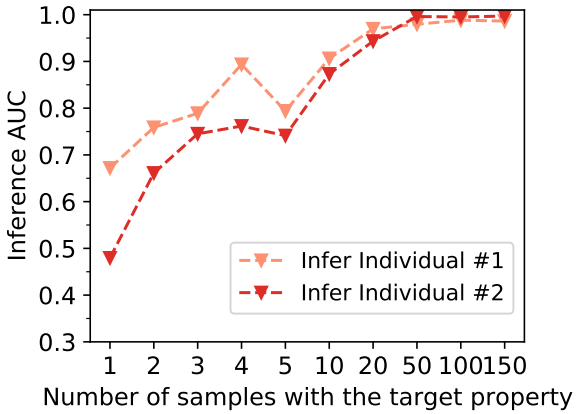

Figure 10. Inference AUC scores when considering multiple properties simultaneously. The inference task is to infer two individuals in the gender recognition setting. The downstream set has 5 000 samples.

In this section, we demonstrate that the attack described in Section 4 can be extended to infer multiple target properties simultaneously. The method is to simply associate different secreting parameters with each property. We conducted experiments using the gender recognition setting with some modifications. The new target properties are the two individuals with the most samples in VGGFace2. In the upstream training, we inject 285 and 257 samples with the property into the upstream training set for the two individuals respectively; we also inject 1 425 samples without the target properties (distribution augmentation in Appendix A.1). For each property, the number of scereting activations is 8 (i.e., $\|\boldsymbol{m}\|_1 = 8$). For the downstream training, the candidate set has 250 samples for each target property and 200 000 samples without the target properties. The rest settings are the same as those in Appendix A.5. The manipulation does not affect the accuracy of the main tasks too much (accuracy drop less than 0.6%). The inferences are also highly successful. Figure 10 summarizes the results of the variance test in discriminating downstream models

trained with a target property from those trained without target properties. The results show that AUC scores exceed 0.85 when $\geq$ 10 out of 5 000 samples are with the property, and are higher than 0.95 when $\geq$ 50 samples have the property.

A.12. Details on Anomaly Detection for Zero-Activation Attack

We consider three common anomaly detection methods: K-means [20], PCA [1] and Spectre [17], where Spectre is the current state-of-the-art. K-means leverages the k-means clustering technique to identify outliers while PCA leverages principal component analysis to identify the outliers. Spectre is an improved version of PCA and works much better than PCA when the attack signature is weak (i.e., the distributional difference is small) [17]. When conducting the anomaly detection, following the common setup in Hayase et al., [17], we filter out $1.5n_t$ (n_t is number of samples with target property) samples, simulating the scenario where the defender does not know the exact n_t, but is able to roughly estimate its value and attempt to find most of them.

Results of Anomaly Detection. We show the detection performance in Figure 11. The results show that conducting anomaly detection can filter out majority of samples with the target property in the downstream set and hence, increase the chance of detecting the manipulation. For example, the Spectre defense can filter out 80% of the samples with the target property in most cases for gender recognition and smile detection, and 60% for age prediction. Anomaly detection effectively finds samples with the target property because the attack mainly focuses on improving attack effectiveness by increasing the distinction between samples with and without property, which makes the attack signature of samples with property much stronger. After finding the possible samples with the target property, the defender can then inspect those samples, and try to find the commonalities and then identify the potential target property. Since the process of finding commonalities in the outliers reported by anomaly detection could be trivial (e.g., most samples have the same property or abnormal activations), we do not perform actual experiments for this part. In Section 8.2, we propose a stealthier design, in which anomaly detection cannot reliably detect samples with the target property and thus cannot find the manipulation.

A.13. Experimental Setup of Stealthier Attacks

In Section 8.3, when preparing upstream models, for $\boldsymbol{m}$, we randomly select 16 activations out of total 1 280 for the gender recognition and also select 196 activations out of total 50 176 for smile detection and age prediction. In practice, the total number of channels in convolutional kernels is not very large and therefore, the defender may still be able to brute-force the manipulated activations if $\boldsymbol{m}$ is chosen only at the channel level. Thus, we also choose to select secreting activations directly for tasks where the first layer of the downstream model is convolutional, which may reduce some of the attack effectiveness. For λ, we prefer a larger value for better inference effectiveness while still evading anomaly detection. Therefore, we performed a linear search starting from 1 and incrementing it by 0.5, and terminating when the attack can no longer evade the mentioned anomaly detection methods. With this strategy, we set $\lambda = 2$ for gender recognition, $\lambda = 1.5$ for smile detection and age detection when the inference targets are senior people and Asian people respectively, and $\lambda = 1$ for smile detection and age detection when the inference targets are specific individuals. α, β, and γ are all set to be 1 in the experiments.

A.14. Adaptive Activation Distribution Checking

The activation distribution checking method needs to be adjusted based on the specific attack method used. Using the modified loss design in Section 8.2, our stealthier attack can automatically evade distribution checking of abnormal zeros, as the secreting activations of samples without target property are also non-zero. Hence, we need to design adaptive detection based on activation distribution checking for the modified attack loss.

With the modified attack loss, we find that activations of samples with the property mixes well with ones without the property, and we fail to find a principled method to distinguish their distribution using the overall activations. Because of the design of the attack loss, the main distributional difference comes from the distributional difference in the secreting activations for samples with and without property (i.e., distributional difference is most significant when we only measure secreting activations), to make progress, we assume the defender will follow a two-stage strategy of first identifying the selected secreting activations and then identifying the distributional difference in the potential secreting activations, with a hope that the distributional difference is significant enough to be detected[1].

Since $\boldsymbol{m}$ is randomly generated with proper number of nonzeros, the brute-force strategy for identifying $\boldsymbol{m}$ is computationally infeasible. For example, for gender recognition experiments, defenders have to try a total of $\binom{1,280}{16}$ ($> 2e36$) forms of $\boldsymbol{m}$ (i.e., $\|\boldsymbol{m}\|_1 = 16$ for a total of 1,280 activations). Therefore, alternatively, we present two methods that attempt to approximately identify $\boldsymbol{m}$ with the hope that the approximately well identified $\hat{\boldsymbol{m}}$ still preserves the significant distributional difference of $\boldsymbol{m}$. The two methods we design are based on the fact that: 1) samples with the target property are rare for practically interesting settings, and 2) in the

[1] We do not exclude the possibility of identifying the distributional difference by still checking the overall distribution, and leave further exploration of such detection strategies as future work.

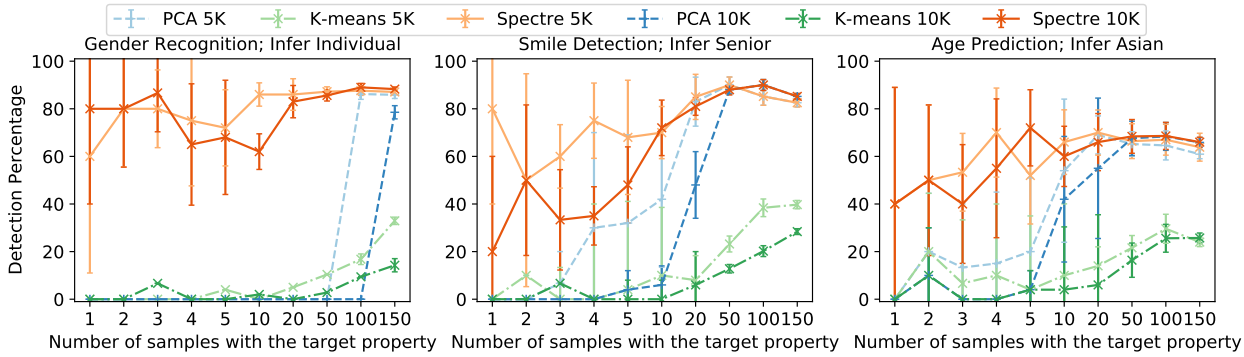

Figure 11. Percentage of samples with the target property detected by the anomaly detection for the zero-activation attack. Similar to [17], we filter out $n \times 1.5$ samples with anomaly detection, where n is the number of samples in downstream training data with the target property. We report the number of samples with the target property filtered out divided by n as the *Detection Percentage*; values are averaged (with standard deviation) over 5 runs of anomaly detection. The '5K' lines report detection results on the settings with 5 000 total samples, while the '10K' lines report for 10 000 total samples.

modified loss design, secreting activations of samples without the property are smaller in magnitude than the ones of samples with the property. Therefore, if we randomly feed inputs to the model, most of the inputs are without property and hence, their corresponding secreting activations should be smaller. With these two principles, we design two detection methods: the first one averages the outputs of each activation for all the fed inputs and treats activations with smaller average values as the potential secreting activations (*average value based detection*); the second approach handles individual input separately and identifies potential secreting activations for each of them, and then returns the intersection for all the potential secreting activations identified (*intersection based detection*). Empirically, we find that both approaches cannot identify the secreting activations well (details are shown below) and hence did not further explore how to check distributional difference on the identified secreting activations in this paper.

Experimental Settings. To evaluate the performance of *average value based detection*, we measure the detection rate, which is the fraction of actual secreting activations in identified potential activations. For the *intersection based method*, since the size of final returned secreting activations can vary (due to intersection over multiple inputs) for different settings, we evaluate the defense performance by reporting their F1-score (viewing actual target as the positive class and others as negative). When running these two detections, we consider an idealized scenario for the defender, where all the randomly sampled inputs are without target property and so, their secreting activations are even smaller for manipulated models and are easier to be detected by the defender.

Specifically, for average value based detection, we choose $n \times 1.5$ activations that have the smallest average values as the identified possible secreting activations (n_{ip}), where n is the number of actual secreting activations ($n = \|\boldsymbol{m}\|_1$). We report the number of identified actual secreting activation (n_{ia}) divided by n as the detection rate. For intersection based detection, the n_{ip} of this method is the number of activations remained after intersection operations, and we cannot precisely control this number. Therefore, only reporting the detection rate like the average value based detection could introduce bias, and we use the F1-score as the metric instead, where the precision is defined as $\frac{n_{ia}}{n_{ip}}$ and the recall is defined as $\frac{n_{ia}}{n}$. And for this detection method, for each sample, we also need to select some activations that have the smallest values as the inputs for conducting the intersection operation. We tried many choices for the number of those activations, and find that choosing $n \times 5$ smallest activations for each sample achieves the best F1-score. In the experiments, we tried to use 100, 500, 1 000, 2 000, 4 000, 8 000, 10 000 samples to generate activations values, separately. For each setting, we repeat each detection 5 times and calculate the average value of the detection rate or F1-score.

Detection Results. Empirically, we find that the two approaches cannot sufficiently identify the secreting activations — the detection rate of secreting activations of the first method is less than 11.3% for gender recognition and is less than 1.5% for smile detection and age prediction for all settings; the F1-score of the secreting activation detection of the second method is less than 0.009 for all settings. In fact, using the second approach, the returned secreting activations are empty sets in most settings, implying the difficulty of identifying the secreting activations by simply checking the magnitude. Overall, the detection performances of both approaches are low and better detection methods are needed for identifying $\boldsymbol{m}$ in the future.

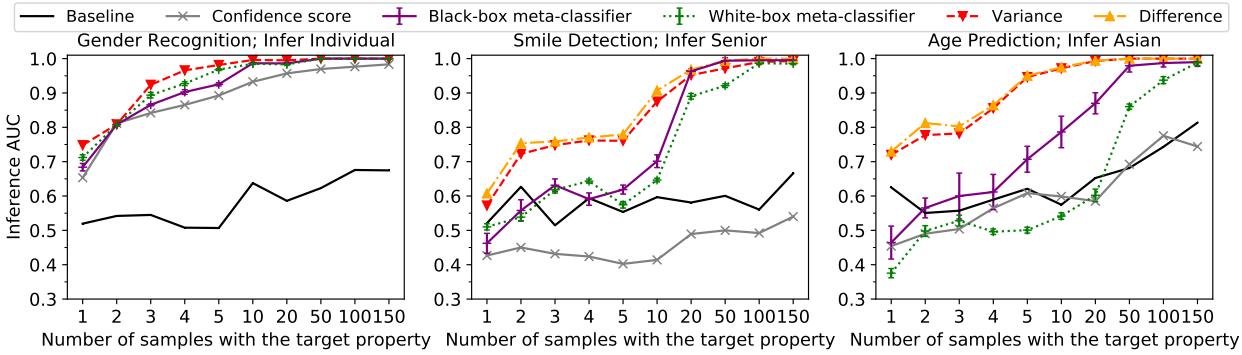

Figure 12. Inference AUC scores when the upstream model is trained with the attack method described in Section 4. Baseline scores (the *baseline* lines) are the maximum of the AUC scores (of the three inference methods) of the baseline experiments in Appendix A.4. The inference of specific individuals for smile detection and age prediction are similarly successfully and found in Figure 15 in the appendix. The downstream training sets have 5 000 samples in the results, and the results for the 10 000 samples are in Figure 1.

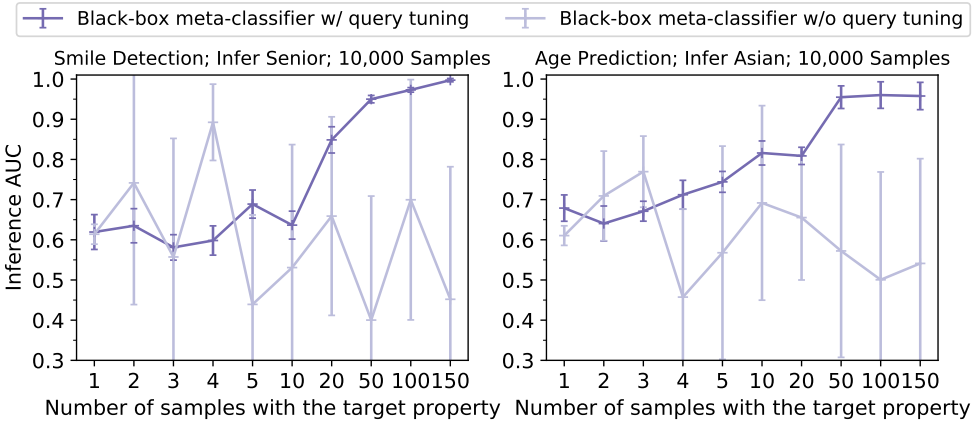

Figure 13. Inference AUC scores of black-box meta-classifiers equipped with and without query tuning. We reuse the upstream and downstream models trained in Figure 1.

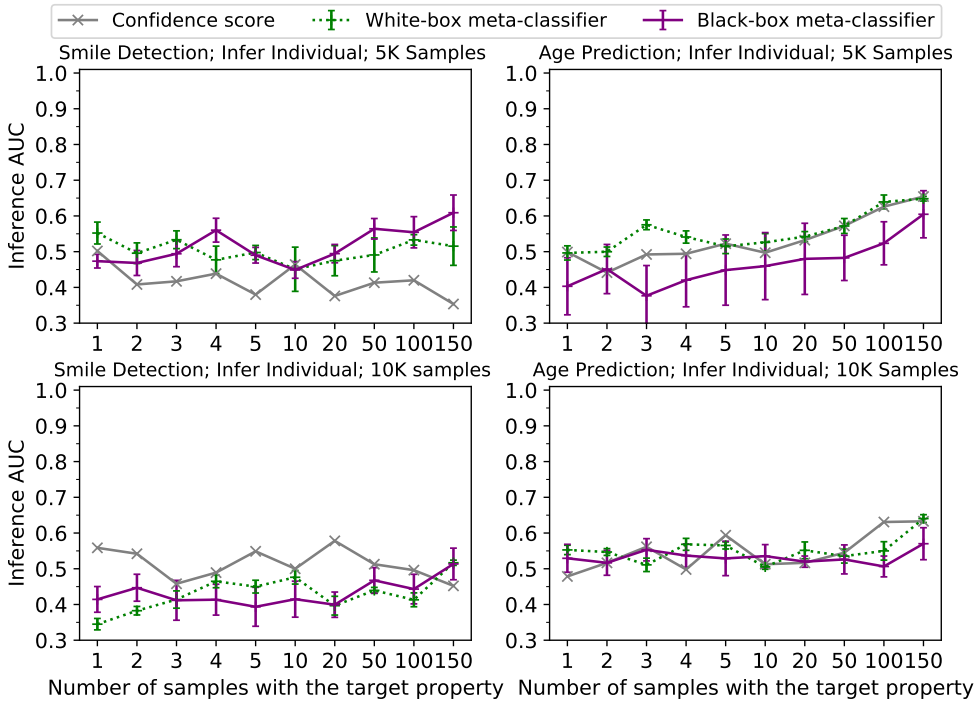

Figure 14. Inference AUC scores when the upstream model is not trained with attack goals. The first and second rows show results when downstream training sets contain 5 000 and 10 000 samples respectively. The inference targets are specific individuals for smile detection and age prediction; the results of other inferences show a similar trend and are found in Figure 3.

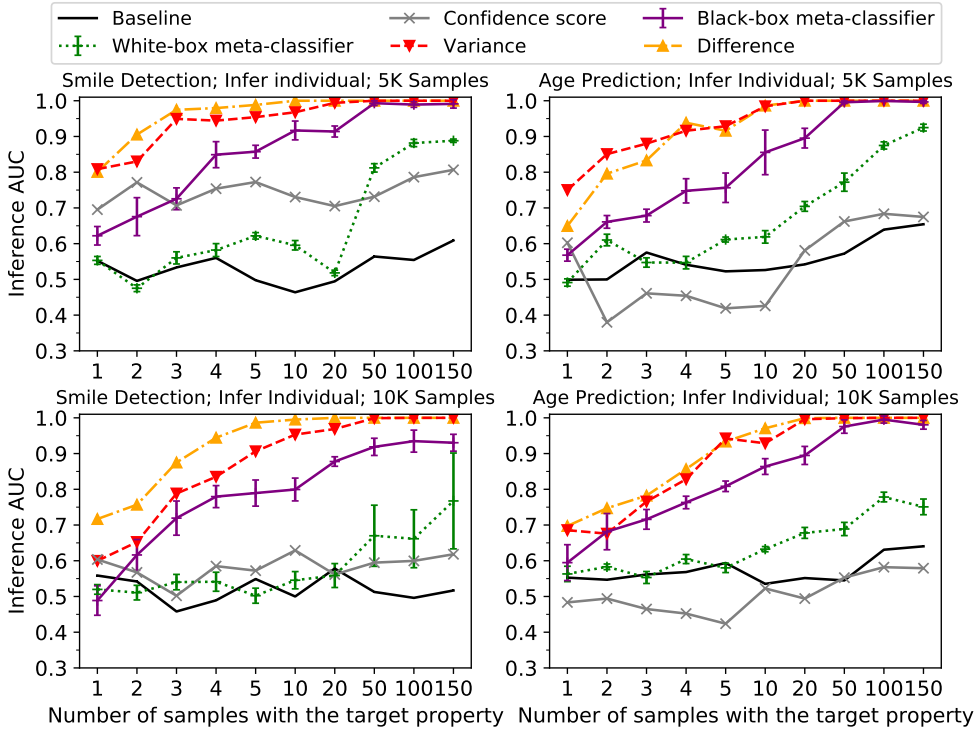

Figure 15. Inference AUC scores when the upstream model is trained with the attack goals described in Section 4. The first and second rows show results when downstream training sets contain 5 000 and 10 000 samples respectively. The inference targets are specific individuals for smile detection and age prediction; the results of other inferences show a similar trend and are found in Figure 1.

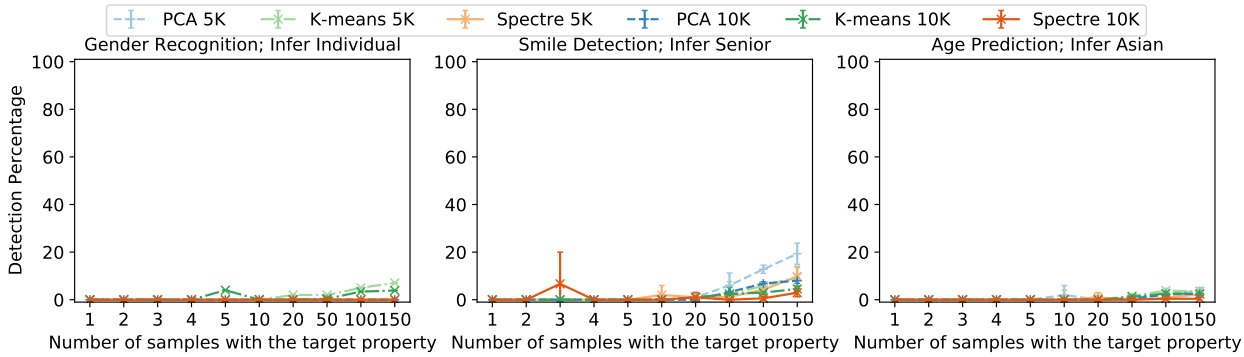

Figure 16. Percentage of samples with the target property detected by the anomaly detection for the stealthier attack. Similar to [17], we filter out $n \times 1.5$ samples with anomaly detection, where n is the number of samples in downstream training data with the target property. We report the number of samples with the target property filtered out divided by n as the *Detection Percentage*; values are averaged (with standard deviation) over 5 runs of anomaly detection. The '5K' lines report detection results on the settings with 5 000 total samples, while the '10K' lines report for 10 000 total samples. Inference targets for smile detection and age prediction are senior people and Asian people respectively; results for the inference of specific individuals follow similar trends (Figure 18).

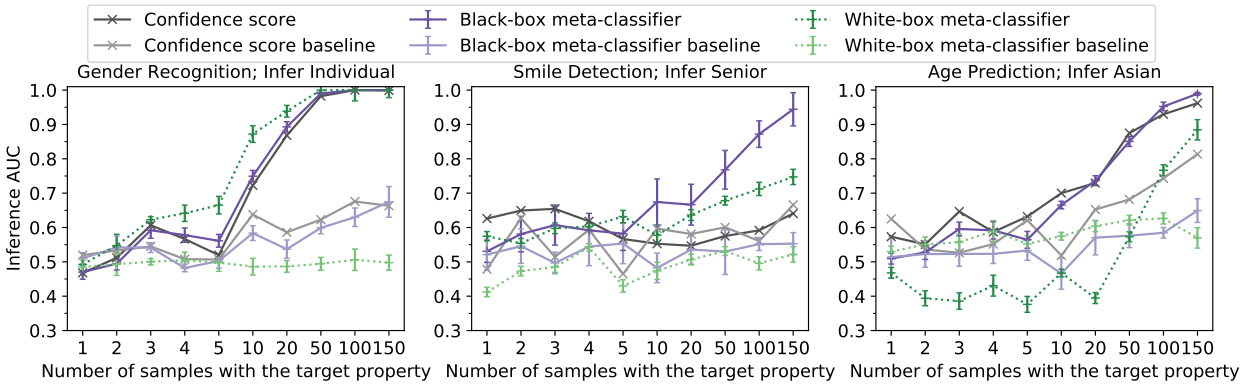

Figure 17. Inference AUC scores of the stealthier design. Since the secreting activations are no longer zero, the inference methods based on difference or variance tests are no longer applicable. Inference targets for the smile detection and age prediction are senior people and Asian people respectively; inference of specific individuals also shows improvement compared to the baseline settings (Figure 19). The downstream training sets have 5 000 samples in the results; results for 10 000 samples show similar trends and are in Figure 2.

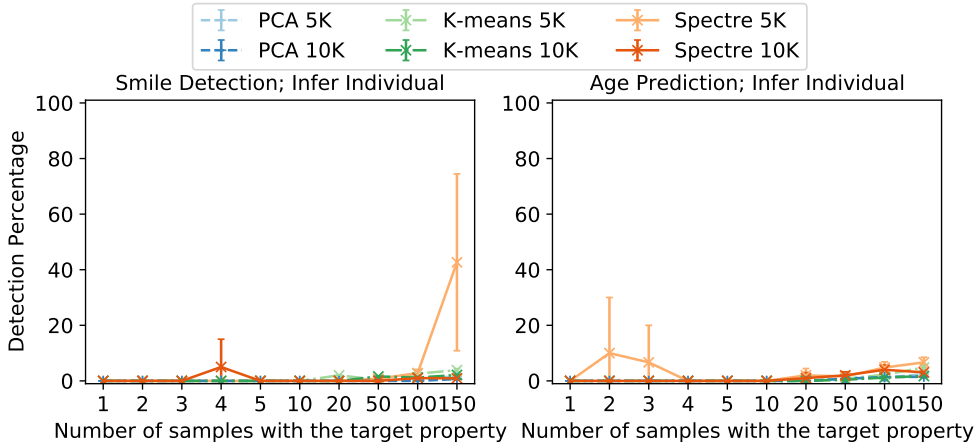

Figure 18. Percentage of samples with the target property detected by anomaly detection for the stealthier attack. The inference targets are specific individuals for smile detection and age prediction; the results of other inferences show a similar trend and are found in Figure 16.

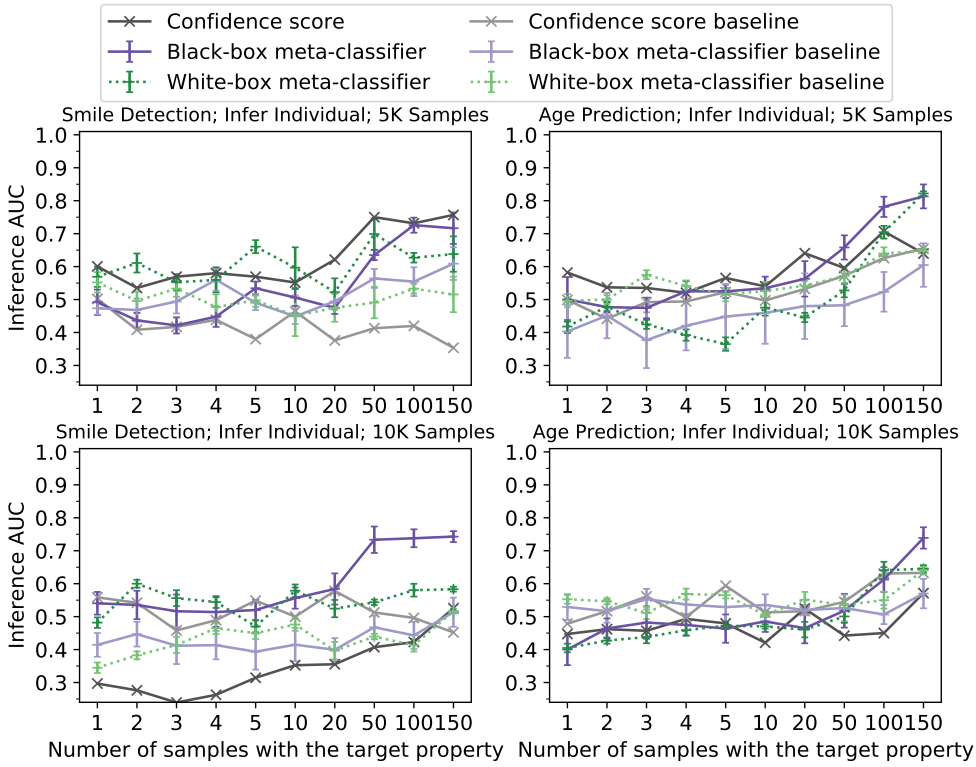

Figure 19. Inference AUC scores of the stealthier attack. The first and second rows show results when downstream training sets contain 5 000 and 10 000 samples respectively. The inference targets are specific individuals for smile detection and age prediction; the results of other inferences show a similar trend and are found in Figure 2.