Property Inference Attacks on Convolutional Neural Networks

Parisot, Mathias P. M.; Pejo, Balazs; Spagnuelo, Dayana

2021

Property Inference Attacks on Convolutional Neural Networks: Influence and Implications of Target Model's Complexity

Mathias P. M. Parisot*, Balázs Pejó†, and Dayana Spagnuelo‡

Abstract.

Machine learning models' goal is to make correct predictions for specific tasks by learning important properties and patterns from data. By doing so, there is a chance that the model learns properties that are unrelated to its primary task. Property Inference Attacks exploit this and aim to infer from a given model (*i.e.*, the target model) properties about the training dataset seemingly unrelated to the model's primary goal. If the training data is sensitive, such an attack could lead to privacy leakage. In this paper, we investigate the influence of the target model's complexity on the accuracy of this type of attack, focusing on convolutional neural network classifiers. We perform attacks on models that are trained on facial images to predict whether someone's mouth is open. Our attacks' goal is to infer whether the training dataset is balanced gender-wise. Our findings reveal that the risk of a privacy breach is present independently of the target model's complexity: for all studied architectures, the attack's accuracy is clearly over the baseline. We discuss the implication of the property inference on personal data in the light of Data Protection Regulations and Guidelines.

1 Introduction

Machine Learning (ML) applications have received much attention over the last decade, mostly due to their vast application range, such as recommendation services, medicine, speech recognition, banking, gaming, driving, and more. It is generally accepted that data plays a vital role in ML models' performance, and that more elaborate models can solve difficult tasks more accurately as they can learn more complex patterns from data. Notwithstanding, besides improving the performance, such ML models introduce privacy issues of the underlying datasets (He et al., 2019).

Training ML models typically requires significant amounts of data, potentially private and sensitive data, and the risk of privacy leakage is not negligible. Suppose a classification model that, once trained, can recognize the appropriate class of a data

*University of Amsterdam, Ansterdam, The Netherlands

†CrySyS Lab, Dept. of Networked Systems and Services, Fac. of Electrical Engineering and Informatics, Budapest Univ. of Technology and Economics, Budapest, Hungary

‡Vrije Universiteit, Amsterdam, The Netherlands

instance by learning mapping patterns between the training dataset and its original set of classes. This mapping is contained within the model parameters, for instance, in a neural network the mapping is encoded in each neuron's weights. As ML models are not exempt from malicious activity, an attacker knowing the trained model parameters could also gain some information about the data it was trained on. This is the rationale of Property Inference Attacks (PIAs) (Ganju et al., 2018; Melis et al., 2019), which aim at uncovering properties of the dataset in which a given model was trained by analyzing the parameters of the model only.

In a world where data has become a commodity—with increased amounts of data generated about people and increased willingness from companies to use such data to gain insights and improve their processes—, regulations come in place to ensure people's fundamental rights to privacy and control over personal data. An example of such is the European General Data Protection Regulation (GDPR)[1]. The GDPR demands that personal data be processed "lawfully, fairly and in a transparent manner", and that principles of data minimization, integrity, and confidentiality be applied[2].

Given the current popularity and the increase in performance of ML applications, it is reasonable to question to which extent the training dataset is vulnerable to privacy attacks and to which extent the processing this data undergoes (e.g., used for ML training) is in line with the current data protection regulations and best practices. In particular, if improving a model's performance is realized by increasing the complexity of the model. As more complex models have more parameters and can retain more information about the training dataset, intuitively one may think that due to this information retention, more complex models could be more sensitive to PIAs as well. In this paper, we study this phenomenon, which—if the model is trained on a dataset containing personal data—could lead to potential privacy leakage.

1.1 Contribution

In this work, we test the influence of a model's complexity on its vulnerability to PIAs. Our setting is processing facial images; thus, we focus on Convolutional Neural Networks (CNNs), the most common model of choice for computer vision tasks. Given that we measure complexity in terms of the number of layers and weights of the model's architecture, *we hypothesize that more complex models are also more sensitive to PIAs.* To uncover whether and how the risk of privacy leaks is influenced by the architecture's complexity of the model, we experiment on nine different CNN architectures. For each, we conduct 30 attacks based on 1500 shadow models.

1.2 Organization

In what follows, we first present works related to ours (Section 2), then we describe the methodology we follow to compose our attack (Section 3). In Section 4 we describe the implementation of our experimental setup, while in Section 5 we present the results of our attack and discuss their meaning. Finally, in Section 6, we present final remarks regarding our findings' legal and ethical implications to data protection, and present future directions for our work.

[1]Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC.

[2]GDPR, Art. 5

2 Related Works

Several security threats are studied regarding machine learning focusing on the basic information security triad: confidentiality, integrity, and availability. For instance (Shumailov et al., 2020) present an ML attack targeting the model's availability. According to (He et al., 2019) the main attack categories for integrity are adversarial and poisoning attacks, while for confidentiality, these are model extraction and model inversion.

Adversarial attacks (Goodfellow et al., 2014; **?**) aim to take advantage of the weaknesses of the target model's classification boundary to craft data instances that are wrongly classified. Poisoning attacks (Mei and Zhu, 2015; **?**) is similar to adversarial attacks as their goal is to influence the prediction of the target model. They do that by polluting the training set with malicious samples. While those are realistic threats to ML models' integrity, they do not pose immediate risk to data privacy. Instead, our work focuses on confidentiality attacks.

Concerning confidentiality (and privacy), the recent survey (Rigaki and Garcia, 2020) gives a comprehensive overview of the subject. Here we only briefly comment on the most relevant works for our purposes, we refer the reader to that work for more details. Model extraction (Tramèr et al., 2016; **?**; **?**) attacks aim at inferring the behavior of the target model to create a substitute model. Model inversion attacks (Fredrikson et al., 2015; Mehnaz et al., 2020) aim at inferring information about the training data, for example, by reconstructing a representative of a particular class of the training set. This type of attack works better when all the instances of a single class represent the same entity. For example, (Fredrikson et al., 2015) train a model to recognize individuals' faces, where the data instances of a given class all represent the same individual. The authors show that it is possible to reconstruct an image of this person. In such a case (Mehnaz et al., 2020) showed that some subgroups are more vulnerable to this type of attack than others.

Depending on the goal of the attacker, we can classify three more attacks under the category of model inversion: membership inference, reconstruction attack, and property inference attack. Membership inference attacks (Truex et al., 2018; Hitaj et al., 2017; Murakonda et al., 2019) aim at determining whether a particular data instance was used for training. This is severe privacy issues when the instance directly maps to an identifiable individual, for instance, a medical records dataset. In contrast, as its name suggests, reconstruction attacks (Zhu and Han, 2020; Li et al., 2019) take this one step further and are capable of recovering both the same training inputs and the corresponding labels.

The last one is PIA, the subject of this study. It aims to infer some hidden properties of the dataset that are independent of any class characteristics, and are therefore not necessarily related to the main classification task. Such properties can be general statistics about the dataset or can reflect biases in the training data.

2.1 Property Inference Attacks

There have been few studies on PIAs with limited results. According to (He et al., 2019), only four papers were published on model inversion attacks, and researchers have not yet entirely determined the vulnerability of neural network architectures to privacy attacks such as PIA. Recent works perform PIAs in a federated learning setting, which allows multiple participants (also called clients by some works) to train a standard model without the need to share data. After each round of training, only the

weights and the gradients are exchanged, while data remains protected on the participants' premises, helping to tackle privacy issues.

The work presented in (Melis et al., 2019) manages to infer properties that hold for a subset of the training data and that are independent of the property the target model aims to predict. Since the attack is performed during the training phase, it requires the model updates that are exchanged between participants. In contrast, the attack we focus on does not require the gradient values after each training round. We also target properties that are true for the whole dataset and not only for a subset.

In (Wang et al., 2019a) three kinds of PIAs are proposed: class sniffing, quantity inference, and whole determination. Class sniffing detects whether a training label is present within a training round. Quantity inference determines how many clients have a given training label in their dataset. The whole determination infers the global proportion of a specific label. All of those attacks extract properties related to classification labels, and therefore to the main classification task. We focus on properties that are, in theory, unrelated to the task of the target model.

Attempts to explore user-level privacy leakage in a federated learning scenario is also subject of recent works (Wang et al., 2019b; Pejó, 2020). They define client-dependent properties to precisely characterize the clients and distinguish them from each other. Both works present an attack that can uncover these differences. In more details, (Wang et al., 2019b) assume a malicious server with access to the individual updates and utilized Generative Adversarial Networks. In contrast, (Pejó, 2020) assume an honest-but-curious setting and recover the participants' quality information via a differential attack without extra computational needs or access to the individual gradient updates.

2.2 Attacks Concerning Model Complexity

A model inversion attack is presented in (Zhang et al., 2020). The authors study and theoretically prove the attack's relation with the model predictive power: more complex models, which should have greater predictive power, should also be more sensitive to model inversion attacks. However, the result of (Zhang et al., 2020) was not proven for PIAs. We also focus more specifically on the complexity of the target model.

Model inversion attacks are also studied in recent works (Geiping et al., 2020). This work analyzes the effects of the target model's architecture on the difficulty of reconstructing input images. The authors investigate attacks on networks with various widths and depths and found that the width has the most significant influence on the reconstruction's quality. Contrary to ours, their study does not consider PIAs and is restricted to federated learning as their attack utilizes the gradient values.

In (Ateniese et al., 2015) the first PIA attack using meta-classifiers is described, this is the methodology of the attack we use in our paper. Their research does not focus on the privacy leakage caused by such an attack but instead on the impact of the training set properties on the model performance. Moreover, they attack models implemented via Support Vector Machines and Hidden Markov Models using a binary tree meta-classifier but do not experiment with deep neural network models.

Finally, an extension of the previous work (Ateniese et al., 2015) is presented by (Ganju et al., 2018). The authors focus on neural networks and notice that a limitation of PIA performance is due to a property of fully connected networks called invariance. They propose two successful strategies to reduce this: converting a neural network to a canonical form, and using a deep set architecture. They use a pre-trained network to generate an embedding, which they feed as input to their target neural network. They

perform the attack using the weights of the fully connected network following the pre-trained one. However, they do not study the influence of the type of layers and the model's architecture on the attack performance, which is the goal of our work.

3 Methodology

This section presents the attack strategy considered for our PIA alongside the assumptions about the target model. To improve readability, we summarize the paper's notations in Table 1.

Table 1: Notations used in the rest of the paper.

Notation	Meaning
M_t	Target model (CNN)
D_t	Training dataset of the target model
P	Property to be inferred
$M_{s_1}, \ldots, M_{s_k}$	Shadow models, mimicking M_t
$W_{s_1}, \ldots, W_{s_k}$	Weights of the shadow models
$D_{s_1}, \ldots, D_{s_k}$	Training dataset of the shadow models
D	The dataset from which D_{s_i} is created
M_a	Attack model to predict P about D_t
D_a	Training dataset of the attack model composed by $W_{s_1}, \ldots, W_{s_k}$

3.1 Threat model

Our target model is a CNN classifier. We assume a training dataset for the classifier, which contains sensitive data according to the definition by the GDPR[3], *e.g.*, data revealing racial or ethnic origin, religious beliefs, or biometric data. We assume an attacker whose goal is to infer general information about the training dataset, such as the proportion of the training data having a property P. This property is unrelated to the main classification task of the model. We assume the attacker can fabricate datasets similar to the original training dataset, *e.g.*, he or she knows from which distribution the original data was created. Moreover, the attacker can manipulate these datasets so that they either contain or not property P. We also assume the attacker has access to the target model and can train a large number of neural networks.

We focus on the white-box setting, where the attacker has access to the target model's full architecture and parameter values. Such information could be obtained in many ways: for instance, it could be shared explicitly as in Federated Learning (Shokri and Shmatikov, 2015). In this scenario, the participants' datasets are hidden from each other (and from the aggregator); however, the participants and the aggregator server have full knowledge of the model and its parameters. Alternatively, when this information is not shared for neural networks, it could still be obtained by creating

[3]GDPR, Art. 9

a substitute model with a similar decision boundary as the target model using a model extraction attack (Papernot et al., 2016).

3.2 Attack

We focus on PIAs whose goal is to extract general information about the target model's training dataset. This information is represented by a property P, which can be true or false. For instance, if the used dataset contains images of faces, P can be defined as *more than 20% of the images within the dataset depict non-white people*.

In this sense, PIA is transformed into a classification problem (for a given model) to determine whether it was trained on a dataset with property P. As such, the attack model can be understood as a meta-classifier as the dataset on which it is trained composed of shadow models, which are themselves classifiers. The attack consists of training k shadow models $(M_{s_1}, \ldots, M_{s_k})$ on k datasets $(D_{s_1}, \ldots, D_{s_k})$ specifically crafted to contain or not the target property P. The training set (D_a) for the attack model (M_a) is composed by the weights $(W_{S_1}, \ldots, W_{S_k})$ of shadow models $(M_{s_1}, \ldots, M_{s_k})$ fabricated with the same architecture as the target model M_t.

The general overview of our attack is described in Figure 1: we train an attack model that takes as input the weights of the target model and outputs the probability the dataset used for training the target model has property P or not. This is based on the baseline PIA presented in (Ganju et al., 2018), which serves our purpose as we do not focus on the PIA itself but rather on the PIA's behavior performed on models with different complexities.

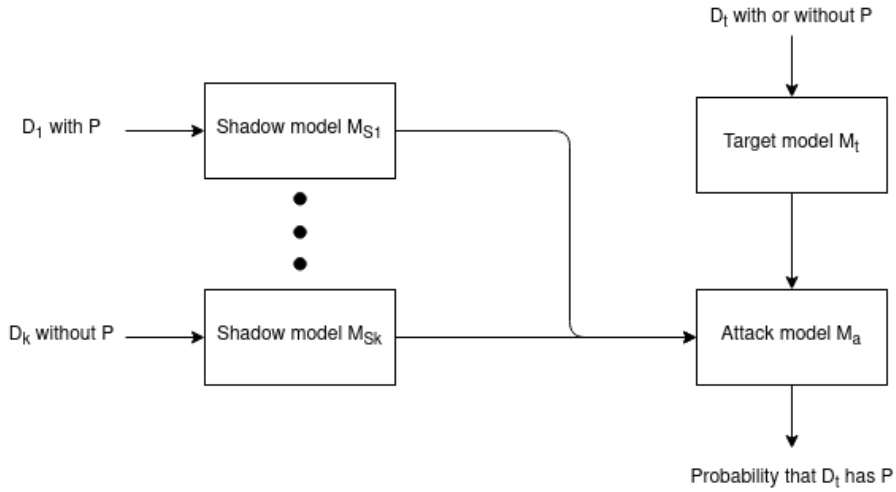


Figure 1: Property Inference Attack using a meta-classifier M_a and datasets of shadow models $\{M_{s_1}, \ldots, M_{s_k}\}$.

4 Experimental Setup

The experiments were performed on a laptop with an Intel i7-8750H (2.20GHz), 8GB RAM, and an Nvidia Quadro P600 GPU. The operating system is Ubuntu 20.04. The

Table 2: Layer-level description of each target & shadow architectures. The detailed description of the parameters in each layer is presented in Table 3.

	Conv. 1	Max-pool	Conv. 2	Max-pool	Conv. 3	Max-pool	FC 1	FC 2	FC 3
A_1	✓	✓	✓	✓	✓	✓	✓	✓	✓
A_2	✓	✓	✓	✓	✓	✓	✓		✓
A_3	✓	✓	✓	✓	✓	✓			✓
A_4	✓	✓	✓	✓			✓	✓	✓
A_5	✓	✓	✓	✓			✓		✓
A_6	✓	✓	✓	✓					✓
A_7	✓	✓					✓	✓	✓
A_8	✓	✓					✓		✓
A_9	✓	✓							✓

shadow models' training and the attack models were both done using Pytorch and are available on a public repository[4].

4.1 Datasets

To train the shadow models for our experiments, we have selected CelebFaces Attributes (CelebA) (Liu et al., 2015) dataset, which contains personal and sensitive information. This is a face attributes dataset containing more than 200.000 face-centered images of 64 by 64 pixels of more than 10.000 celebrities. The images are labeled using 40 physical attributes such as hair color, smiling, and wearing a hat. Our shadow models and the target model are trained to detect whether the person appears with their mouth open in a given photo. Hence, we use the *Mouth_Open* attribute of the dataset.

Although this might seem like an irrelevant classification task, at the time of execution of this work, the world is undergoing a pandemic, and mouth covering masks is one of humanity's currently available weapons against the SARS-CoV-2 virus (Eikenberry et al., 2020). This task can be related to automated mask detection[5], especially since, in many places, their use is compulsory.

4.2 Shadow Models

The shadow models $\{M_{s_1}, \ldots, M_{s_k}\}$ are trained to mimic the target model M_t, *i.e.*, to differentiate between images of persons with and without their mouth open. However, the attacker's goal is to infer whether the training set of a given model was composed of an unbalanced number of images of males. This targeted property concerns biometric data and is classified as sensitive data according to the GDPR. In the real world, although it is not common, it is still possible to have gender-unbalanced generations in some societies: in the next 15 years in large parts of China and India, it is estimated that there will be around 20% excess of young men (Hesketh and Min, 2012), which corresponds to 60%-40% division. Consequently, we set our thresholds to be at least double this gap: *P is true when the model's training set is composed of 70% or more images containing males*. It is important to note that P is not related to the model's

[4]The exact source will be shared after the blind review process to maintain anonymity.

[5]`http://tinyurl.com/2a8ewzvl`

classification task and that the target model does not use, at any time during training, the gender attribute.

The shadow models $\{M_{s_1}, \ldots, M_{s_k}\}$ have the same architecture as the targeted model and are trained to a reasonable level of accuracy to mimic the target model's behavior. We set the lowest acceptable accuracy to be 85% (on the mouth open classification task) when the baseline distribution of the dataset is 51.7%. For the attack to be effective, many shadow models have to be trained, as the attack model's inputs are the weights of the shadow models. Hence, for a specific target model architecture, we train 1800 shadow models. Since the computational cost of training many shadow models is significant, we decide not to use all images of *CelebA*. Rather, for each shadow dataset $\{D_{s_1}, \ldots, D_{s_k}\}$, we use only 2000 randomly selected images. For half of the shadow models (i.e., 900 times) these 2000 images were selected to have property P, while the remaining 900 datasets do not. In detail, the exact proportion of males for each dataset was randomly taken from a uniform distribution either above or below 70%, respectively. It is important to note that while each shadow model is trained using only 2000 images, they perform with at least 85% accuracy on the whole test set of CelebA; therefore, they do not overfit to their small training set.

We experimented on target model's (and consequently the shadow models') architectures composed of up to 9 layers, each of three kinds: convolution layers, pooling layers, and fully connected layers. We trained 9 architectures $(A_1, \ldots, A_9)$ which are presented in Table 2, while the description of each layer is presented in Table 3. The models take as input 64 by 64 RGB face images and output each picture's probability of representing a person with mouth open. Every architecture comprises 1-3 convolution layers, each followed by a max-pooling layer with a ReLU activation and 1-3 fully connected layers with a ReLU activation. The shadow models are trained for 50 epochs using the Mean Squared Error loss and the Adam optimizer with a learning rate of 0.001 and without decay or regularization.

Table 3: Description of the different layers used in the target & shadow architectures.

Layer	Description
Convolution 1	6 filters 5x5
Max-pool	2x2, ReLU
Convolution 2	16 filters 5x5
Max-pool	2x2, ReLU
Convolution 3	32 filters 5x5
Max-pool	2x2, ReLU
Fully-Connected 1	120 neurons, ReLU
Fully-Connected 2	84 neurons, ReLU
Fully-Connected 3	1 neuron

4.3 Attack Model and Evaluation

The attack model classifies shadow models on whether they were trained on a dataset with or without the property P. The dataset used for the attack is composed of the 1800 shadow models and is split into training (1500 shadow models), validation (100 models), and test sets (200 models). The training algorithm is presented in Algorithm 1. The attack model is a simple multi-layer perceptron tuned using the validation set

and evaluated on the test set.

Algorithm 1 Attack model training

procedure TRAIN_ATTACK(D, k)
  D dataset of images, k number of shadow models to train
  let: D_{s_i} be a subset of D
  let: D_a be the dataset used to train the attack model
  let: P_{s_i} be a boolean value determining whether P is True on D_{s_i}
  $D_a \leftarrow \{\}$
  for $i \leftarrow 1, k$ **do**
    $D_{s_i} \leftarrow$ sample subset of D
    $P_{s_i} \leftarrow$ evaluate P on D_{s_i}
    $M_{s_i} \leftarrow train(D_{s_i})$
    $W_{s_i} \leftarrow getWeights(M_{s_i})$
    $D_a \leftarrow D_a \cup \{(W_{s_i}, P_{s_i})\}$
  end for
  $M_a \leftarrow train(D_a)$
  return M_a
end procedure

We tuned the attack model by performing a grid search over the following hyper-parameters: learning rate, loss function, batch size, optimizer, the activation function of the first layer of the attack model. The hyper-parameters values we used are presented in Table 4. We trained six attack models for the 9 model architectures for each of the combinations of parameters during ten epochs. We selected the best parameters by combining the largest median accuracy over the 9 model architectures.

Table 4: Attack model hyper-parameter tuning with the optimal ones marked in bold.

Parameter	Values
Learning rate	**0.005**; 0.001; 0.0005
Loss function	**MSE**; L1-loss
Batch size	16; **32**; 64
Optimizer	SGD; **Adam**
Input layer act. func.	sigmoid; **ReLU**; tanh

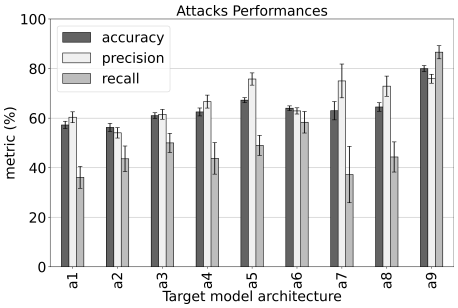
After tuning, the selected attack model architecture is presented in Table 5. The attack model's inputs are the flattened weights of the model it is trying to classify as having or not the property P. Therefore, a target model architecture with a larger number of parameters induces a more comprehensive input layer for the attack model. The attack model comprises two fully connected layers, which we trained 30 times for each shadow model architecture. We presented the average performances after 20 epoch when the training was done using the Mean Squared Error loss function and the Adam optimizer with a learning rate of 0.005 and without decay or regularization.

Table 5: Architectures of the attack model. FC1 is the input layer and FC2 the output layer.

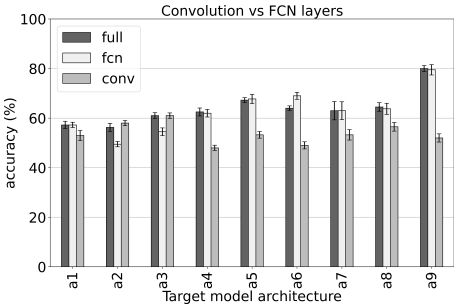
Name	Description
Fully-Connected 1	10 neurons, ReLU
Fully-Connected 2	1 neuron

5 Results and Discussions

Figure 2 summarizes the performance of our attack. In detail, Figure 2a presents the accuracy of the attacks on each target model architecture, which varies between $56\% - 80\%$ depending on the architecture. These results confirm the findings of (Ganju et al., 2018) that the target models do learn information unrelated to the task they were trained to learn.



(a) Attack's accuracy, precision, and recall using weights from entire target model.



(b) Attack's accuracy using weights from: entire target model (full), only fully connected layers (fcn), and only convolution layers (conv).

Figure 2: Attack's performance on each architecture. Bars correspond to the median of 30 attacks, and error bars to $\pm$ the standard deviation.

We create as many shadow models presenting the property P as ones not presenting it in our setup. Therefore, the expected baseline is 50% accuracy. Most of the architectures have less than 67% attack accuracy, so the real privacy threat is low. However, the attack model was tuned to achieve good performance across the nine shadow model architectures. Fine-tuning the attack model for a specific architecture should improve the attack accuracy. Moreover, (Ganju et al., 2018) have shown that it is possible to significantly increase the accuracy of this attack using representations that are invariant to node permutations. They managed to get almost perfect accuracy making the attack usable in a real-world setting.

We performed PIAs on distinctive neural network architectures with different amounts and types of layers. As convolution layers and fully connected ones play different roles in a CNN, we also tested whether the type of used layers impacts the attack's accuracy. Thus we conducted three additional PIAs on each of the architectures presented in Table 2: 1) using all the weights of the shadow model; 2) using only the weights of the convolution layers, and 3) using only the weights of the fully connected layers. Figure 2b presents the accuracy of the three attacks. For most target model architectures, the

PIA using only the fully connected weights performs as well, and sometimes better, as the PIA using the weights from both types of layers. Consequently, the information leaked by a CNN seems to be contained in the fully connected part of the network.

One of this study's goals is to establish whether there is a relationship between model complexity (which we define by the number of parameters) and its sensibility to PIAs defined as the accuracy of the attack. Figure 3 summarizes this relationship. Our results do not directly support this claim.

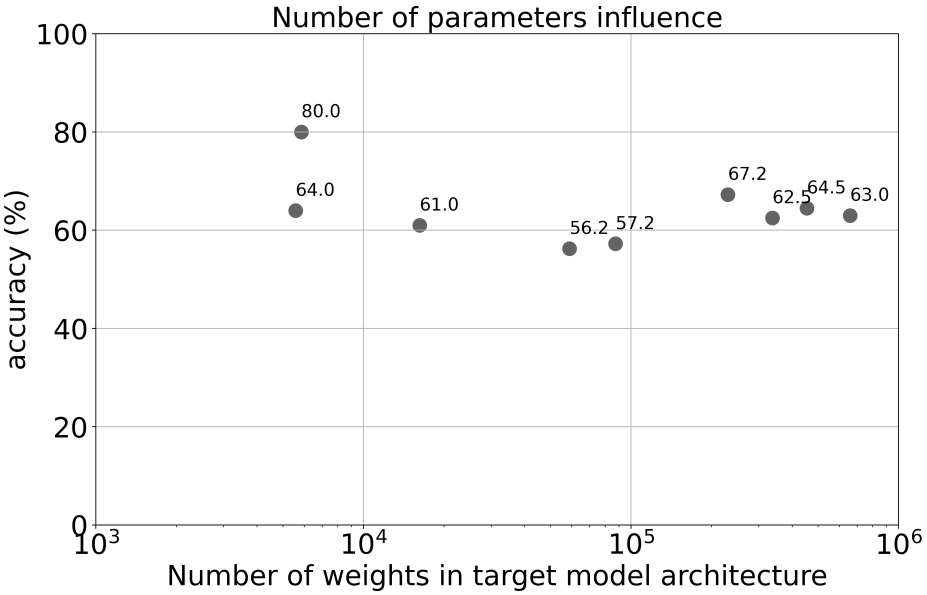


Figure 3: Influence of the complexity of the target model (express as the number of parameters) on the accuracy of the attacks on each architecture. Each dot corresponds to the median of 30 attacks.

6 Conclusion

This work presents an attack that tries to determine if a given dataset (of faces) used to train a CNN model (to determine if a mouth is open in a picture) has some property P, in our case, whether the dataset was unbalanced gender-wise. We conducted several experiments to uncover the relationship between target model complexity and privacy leakage (aka PIA's accuracy), and we find no significant correlation between the two.

Due to the nature of PIAs, our work has an explicit limitation: it is tailored to one specific property of the dataset. Our attack can be adapted to other properties P as well, as long as the attacker can fabricate datasets containing or not the given property. It is possible, for instance, to infer whether a training set contains pictures of a given person. In this case, when the instances of the same class represent the same entity (same person), a PIA is transformed into a membership inference attack which is not only focused on a specific data instance but on whether the whole dataset contains a particular class of image (a particular person). This type of membership inference attack presents a distinct type of threat. Regular membership inference attacks rely on the attacker to possess the same data instance present in the dataset. For instance, if the

dataset is composed of portrait images, the attacker must possess the same image of a given person that belongs to the dataset. It may not be realistic to assume that attackers will get hold of the very same image. However, it is more reasonable to expect an attacker can produce a dataset of different targeted persons' pictures. The attack then consists of creating shadow datasets with and without pictures of the targeted person and using them to train shadow models and perform a PIA.

6.1 Data Protection Implications

Our experimental attack highlights a surplus of personal and sensitive data for training the classification models (attacks' accuracy are equal or higher than 56%). In the context of the current data protection regulation (*i.e.*, GDPR), this can be framed as a violation of the *data minimization* principle of data processing, which explicitly requests that processing of personal data be "adequate, relevant and limited to what is necessary concerning the purposes for which they are processed"[6]. In this paper, the PIA demonstrates that for the task of classifying whether someone's mouth is open, there is a leak of the gender balance in the training dataset. This can be attributed to the fact that the dataset comprises full faces instead of only mouth images, which is arguably the only data indeed necessary for the task.

An intuitive solution for the surplus of data problem would be to crop images to contain only the area relevant for the classification task. In our example, a preliminary step identifies mouths in the images and eliminates the remaining of the face. However, this alone could prove insufficient in the context of the attack we present in this paper. As empirically demonstrated by Pew Research Center[7], gender traits can be encoded in many (unexpected) areas of a facial image. Recent works explore alternative techniques that could help achieve data minimization in images, such as feature anonymization (Kim and Yang, 2020), and gender obfuscation through morphing (Wang, 2020). A possible future research direction would be to rerun our experiment with models trained on datasets using those techniques and test their impact on PIAs.

The GDPR also demands that people be informed, among others, of the envisioned consequences of data processing, which leads to automated decision-making[8]. This request is part of the concept of transparency, which is essential to ensure that people can exercise their rights under the GDPR (Demetzou, 2018). Our PIAs brings to light one of the hidden threats to people's data due to such processing. However, it is reasonable to assume other types of attacks may uncover different threats. While informing of possible risks seems fundamental to grant people's right to privacy, thoroughly doing so may require disproportional technical efforts from the data controller's side. For instance, requiring controllers to run experiments similar to the one presented in this paper. Some works in the literature seem to agree with the vision that risks and consequences of "undesired outcomes" from machine learning models are of interest to people whose data is processed (Ras et al., 2018). Nevertheless, the question that remains is to what extent should a data controller explore the possible risks to *fairly* inform people of *envisioned* consequences. Answering this question is a task we leave for future works.

More conservative school of thought argues that machine learning models— specifically the ones for which privacy attacks are possible —might be seen as a type of personal data themselves (Veale et al., 2018). The authors argue that model inversion

[6]GDPR, Art. 5.1

[7]https://www.pewresearch.org/interactives/how-does-a-computer-see-gender/

[8]GDPR, Art. 13.2(f)

(of which PIA is a type, and aims at estimating the training data from a publicly available model), despite some level of uncertainty, leads to recovering of data that will be more accurate than simply guessing characteristics of the original training dataset. A comparison is made with 'pseudonymization'[9], which aims at transforming data in a way that can no longer identify a person without the use of additional data[10]. The authors present a legal precedent that shows this data need not be all in possession of the same entity for the re-identification to be considered a privacy breach (*Breyer* (ECLI:EU:C:2016:779) as cited in (Veale et al., 2018)). In the same way, PIAs uncovering new data about the data set would also characterize a privacy breach, even if no personal data is made available by the data controllers themselves. This is turn triggers a series of rights for the data subjects (people whose data are in the dataset), and obligations for the data controller (organization training and using the model), among those the one of complying with the security and data protection by design principles. The conclusion is that sharing or making available models for which privacy attacks are possible, if done with no legal basis (which is not needed for anonymous or non-personal data), would be characterize a violation of such principles and a privacy breach.

Even if such arguments might feel unrealistic at the moment, we draw attention to guidelines recently published by public authorities suggesting the best practices for for Data Protection in Artificial Intelligence (AI). They evidence that a requirement for actively thinking about the risks of PIAs and other types of attacks to machine learning models is in the horizon. The guidelines on AI and data protection by the Council of Europe call for *algorithm vigilance* and suggest legislators and policy makers to require "prior assessment of the impact of data processing on human rights and fundamental freedoms, and vigilance on the potential adverse effects and consequences of AI applications" (Section III.2 in (Council of Europe, 2019)). Other similar guidelines also emphasize the role of assessing the expected impacts and risks (Data Protection Impact Assessment) for the data subject in the beginning of an AI project (Commission Nationale de l'Informatique et des Libertés (CNIL) et al., 2018; **?**).

Concluding remarks. In this work, we present an experimental setting to test the influence and implications of the target model's complexity in the accuracy of Property Inference Attacks (PIAs). Although our findings did not support our initial hypothesis that more complex models would intrinsically learn more information from the training dataset and hence be more sensitive to PIAs, they reveal a surplus of personal information used in the training stage of CNN models. The implications of our work is shown to have an impact on the rights and obligations with respect to Data Protection Regulations and Guidelines.

REFERENCES

Ateniese, G., Mancini, L. V., Spognardi, A., Villani, A., Vitali, D., and Felici, G. (2015). 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.

Commission Nationale de l'Informatique et des Libertés (CNIL), European Data Protection Supervisor (EDPS), and Garante per la protezione dei dati personali (2018). Declaration on ethics and data protection in artificial intelligence.

[9] A note on terminology: the notion of 'psuedonymization' as described by the GDPR resembles what works from the discipline of computer science refer to as 'anonymization'.

[10] GDPR, Art. 4.5

http://globalprivacyassembly.org/wp-content/uploads/2018/10/20180922_ICDPPC-40th_AI-Declaration_ADOPTED.pdf. Last accessed 25 February 2021.

Council of Europe (2019). Guidelines on artificial intelligence and data protection. https://rm.coe.int/guidelines-on-artificial-intelligence-and-data-protection/168091f9d8. Last accessed 25 February 2021.

Demetzou, K. (2018). GDPR and the Concept of Risk. In *IFIP International Summer School on Privacy and Identity Management*, pages 137–154. Springer.

Eikenberry, S. E., Mancuso, M., Iboi, E., Phan, T., Eikenberry, K., Kuang, Y., Kostelich, E., and Gumel, A. B. (2020). To mask or not to mask: Modeling the potential for face mask use by the general public to curtail the covid-19 pandemic. *Infectious Disease Modelling*.

Fredrikson, M., Jha, S., and Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333.

Ganju, K., Wang, Q., Yang, W., Gunter, C. A., and Borisov, N. (2018). Property inference attacks on fully connected neural networks using permutation invariant representations. pages 619–633.

Geiping, J., Bauermeister, H., Dröge, H., and Moeller, M. (2020). Inverting gradients–how easy is it to break privacy in federated learning? *arXiv preprint arXiv:2003.14053*.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

He, Y., Meng, G., Chen, K., Hu, X., and He, J. (2019). Towards privacy and security of deep learning systems: a survey. *arXiv preprint arXiv:1911.12562*.

Hesketh, T. and Min, J. M. (2012). The effects of artificial gender imbalance: Science & society series on sex and science. *EMBO reports*, 13(6):487–492.

Hitaj, B., Ateniese, G., and Perez-Cruz, F. (2017). Deep models under the gan: information leakage from collaborative deep learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 603–618.

Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C., and Li, B. (2018). Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 19–35. IEEE.

Kim, T. and Yang, J. (2020). Selective feature anonymization for privacy-preserving image data publishing. *Electronics*, 9(5):874.

Li, Z., Huang, Z., Chen, C., and Hong, C. (2019). Quantification of the leakage in federated learning. *arXiv preprint arXiv:1910.05467*.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Mehnaz, S., Li, N., and Bertino, E. (2020). Black-box model inversion attribute inference attacks on classification models. *arXiv preprint arXiv:2012.03404*.

Mei, S. and Zhu, X. (2015). Using machine teaching to identify optimal training-set attacks on machine learners. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Melis, L., Song, C., De Cristofaro, E., and Shmatikov, V. (2019). Exploiting unintended feature leakage in collaborative learning. pages 691–706.

Murakonda, S. K., Shokri, R., and Theodorakopoulos, G. (2019). Ultimate power of inference attacks: Privacy risks of learning high-dimensional graphical models. *arXiv preprint arXiv:1905.12774*.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., and Swami, A. (2016). Practical black-box attacks against deep learning systems using adversarial examples. *arXiv preprint arXiv:1602.02697*, 1(2):3.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., and Swami, A. (2017). Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519.

Pejó, B. (2020). The good, the bad, and the ugly: Quality inference in federated learning. *arXiv preprint arXiv:2007.06236*.

Ras, G., van Gerven, M., and Haselager, P. (2018). Explanation methods in deep learning: Users, values, concerns and challenges. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*, pages 19–36. Springer.

Rigaki, M. and Garcia, S. (2020). A survey of privacy attacks in machine learning. *arXiv preprint arXiv:2007.07646*.

Shokri, R. and Shmatikov, V. (2015). Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1310–1321.

Shumailov, I., Zhao, Y., Bates, D., Papernot, N., Mullins, R., and Anderson, R. (2020). Sponge examples: Energy-latency attacks on neural networks. *arXiv preprint arXiv:2006.03463*.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

The Norwegian Data Protection Authority (Datatilsynet) (2018). Artificial intelligence and privacy. `https://www.datatilsynet.no/globalassets/global/english/ai-and-privacy.pdf`. Last accessed 25 February 2021.

Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., and Ristenpart, T. (2016). Stealing machine learning models via prediction apis. In *25th {USENIX} Security Symposium ({USENIX} Security 16)*, pages 601–618.

Truex, S., Liu, L., Gursoy, M. E., Yu, L., and Wei, W. (2018). Towards demystifying membership inference attacks. *arXiv preprint arXiv:1807.09173*.

Veale, M., Binns, R., and Edwards, L. (2018). Algorithms that remember: model inversion attacks and data protection law. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133):20180083.

Wang, B. and Gong, N. Z. (2018). Stealing hyperparameters in machine learning. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 36–52. IEEE.

Wang, L., Xu, S., Wang, X., and Zhu, Q. (2019a). Eavesdrop the composition proportion of training labels in federated learning. *arXiv preprint arXiv:1910.06044*.

Wang, S. (2020). Gender obfuscation through face morphing. Master's thesis, University of Twente.

Wang, Z., Song, M., Zhang, Z., Song, Y., Wang, Q., and Qi, H. (2019b). Beyond inferring class representatives: User-level privacy leakage from federated learning. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pages 2512–2520. IEEE.

Zhang, Y., Jia, R., Pei, H., Wang, W., Li, B., and Song, D. (2020). The secret revealer: Generative model-inversion attacks against deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 253–261.

Zhu, L. and Han, S. (2020). Deep leakage from gradients. In *Federated Learning*, pages 17–31. Springer.