Adaptive Image Anonymization in the Context of Image Classification with Neural Networks

Nadiya Shvai[1]* Arcadi Llanza Carmona[1,2] Amir Nakib[1,2]
[1] Cyclope.ai, Paris, France
[2] University Paris Est Créteil, Laboratoire LISSI, Paris, France
nadiya.shvai@cyclope.ai, {arcadi.llanza-carmona, nakib}@u-pec.fr

Abstract

Deep learning based methods have become the de-facto standard for various computer vision tasks. Nevertheless, they have repeatedly shown their vulnerability to various form of input perturbations such as pixels modification, region anonymization, etc. which are closely related to the adversarial attacks. This research particularly addresses the case of image anonymization, which is significantly important to preserve privacy and hence to secure digitized form of personal information from being exposed and potentially misused by different services that have captured it for various purposes. However, applying anonymization causes the classifier to provide different class decisions before and after applying it and therefore reduces the classifier's reliability and usability. In order to achieve a robust solution to this problem we propose a novel anonymization procedure that allows the existing classifiers to become class decision invariant on the anonymized images without any modification requires to apply on the classification models. We conduct numerous experiments on the popular ImageNet benchmark as well as on a large scale industrial toll classification problem's dataset. Obtained results confirm the efficiency and effectiveness of the proposed method as it obtained 0% rate of class decision change for both datasets compared to 15.95% on ImageNet and 0.18% on toll dataset obtained by applying the naïve anonymization approaches. Moreover, it has shown a great potential to be applied to similar problems from different domains.

1. Introduction

Deep learning models require, in general case, a vast load of data for training, control of performance and analysis of the models. For various reasons such as cost of labeling or necessity to do historic analysis, it might be necessary to keep the data for a longer period of time. For numerous automatic system based service providers such as surveillance, security, toll classification, these data often concerns personal information including identity via face, location with car license plates, etc. Then, saving these information certainly raises the concern on personal privacy and hence on securing the digitized form of personal information from being exposed and potentially misused. As a consequence, multiple privacy protection laws such as General Data Protection Regulation (GDPR) [19] in European Union, California Consumer Privacy Act (CCPA) in California, USA, China Cybersecurity Law (CSL) in China, amended Act on the Protection of Personal Information (APPI) in Japan etc. impose severe limitations on data operations in order to protect private data of the end customer. Indeed, one potential solution of the data saving necessity for the services vs the restrictions imposed by the privacy preservation laws is to remove the sensitive information (*e.g.,* face, license plate) from the data via *anonymization*. Particularly, for the image related tasks anonymization means removing sensitive information from the image by modifying its corresponding part (for example, by blurring). *This research considers the problem within the context of large scale image classification.*

Recent research demonstrates that naïvely applying anonymization effects the performance of the deep convolutional neural network (CNN) based classifiers [10]. Particularly, anonymization causes the classifier to provide different class decisions before and after applying it and hence reduces the classifier's reliability and usability. This creates dual inference pipeline: classifier vs. anonymizer + classifier, where the invariance of the predicted class is not *a priori* guaranteed. Fig. 1 illustrates this problem on an image from ImageNet, where the classifier predicts a different class once the image region is anonymized.

Intuitively, the dual inference pipeline has various immediate solutions such as: (1st) developing the classifiers directly with the anonymized data and (2nd) improve the classifier by training it with both original and anonymized data. Indeed, for any offline (non real-time) or time-

*Corresponding author

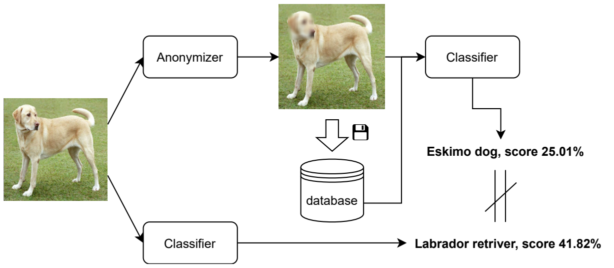

Figure 1. Dual inference pipeline: classifier vs. classifier + anonymizer. The outcomes of such pipeline are not necessary the same.

unconstrained applications the 1st approach is a potential solution given the possibility to work directly with the anonymized data. However for the real-time applications it is not feasible to apply on-the-fly anonymization before applying a deep learning model due to the constraints on operations time and available resources. Next, the second solution has been examined in [10], which demonstrated that it is unable produce acceptable results. Furthermore, an in-depth analysis of this problem by recent research [10] reveals that the dual inference problem is mainly caused due to the vulnerable samples and hence the robust classifier design appears as the best solution. Developing a robust classifier is closely related to numerous concurrent approaches for improving classifier's reliability against the adversarial attacks. Unfortunately, despite the all recent improvements the problem is yet to achieve its desire performance, which is 0% rate of decision change.

The above limitations of the possible and existing approaches highly motivate us to rethink the problem from different aspects to identify novel and effective solution. Therefore, we address it from the aspect of modifying the perturbations introduced by the anonymizer in such a way that they do not significantly change the output of the classifier model (here the term *significantly* is defined by specific problem conditions). Consequently, we propose *a novel anonymization procedure that allows the existing classifiers to become class decision invariant on dual inference pipeline* without applying any modification on the classification models. Particularly, using small changes in the area perturbed by anonymizer model, we "correct" the image in order to preserve the predicted class of the original image. In order to find changes necessary to apply necessary modifications we use the gradient descent method, which in its base is the same idea as used for gradient adversarial attack. The proposed method provides numerous advantages:

- it allows to keep the classification model intact, which can be especially interesting in the situation of legacy model, or costly model training and validation processes;
- the method is simple in implementation;
- there is no dependency on the specific neural network architecture, or even specific task, although in our work we discuss only image classification;
- proposed method has low number of parameters;
- it can work directly "out-of-the-box" as we show in the experiment section where we apply it with different CNN architectures without specific tuning of method parameters.

In order to validate the effectiveness of the proposed approach, we conduct experiments with two different datasets and associated image classification problems: ImageNet and a proprietary toll vehicle classification dataset. The first set of experiments on the open dataset aims to provide reproducible results and simulates a rather difficult problem in order to explore method limits. The second set of experiments illustrates the method's performance on a real-world application (from which this work originated). Obtained results from both datasets confirm the effectiveness of the proposed method as it obtained 0% rate of class decision change for both compared to 15.95% on ImageNet and 0.18% on toll dataset obtained by applying the naïve anonymization approaches. Our contributions in this research can be summarized as follows:

- we propose a novel solution for an important and contemporary problem, introduced as the dual inference pipeline;
- we provide a novel view of the problem to be experimented with the publicly available popular ImageNet dataset.
- we achieved the desire and optimal performance, which is 0% rate of decision change.

The paper is organized as follows: section 2 discusses related work, in section 3 we set the problem, describe the proposed method and give more details on experiments settings. Section 4 contains the experiments results and observations made throughout conducting those. Finally, section 5 concludes the paper.

2. Related work

When establishing the related work, we can do so both through the *problem domain* (image anonymization) or the *proposed solution* (gradient-based image modification). Finally, considering the complete pipeline of anonymizer and classifier and the issue of its *invariance*, we touch the topic of classifier robustness which is an alternative solution to the posed problem.

2.1. Image anonymization

This work falls under the broad category of research aiming to *modify the input for the purpose of privacy preserving issues.* The main goal is to be compliant with the different privacy related regulations such as GDPR, CCPA etc. These regulations in particular aim to preserve the private part of data used to build machine learning based services. In the context of computer vision, an easy way to implement privacy protection would be to modify a part of an image to hide private information such as faces or license plates.

Generally speaking, we can think about anonymization as a type of input mapping that erases or encrypts some part of information while leaving the rest of it intact. To solve the problem of finding such a mapping correctly a measure of information loss and preservation needs to be established. This measure is problem dependant and can be subjective. In our case the information preservation is ensured by a/ the assumption of targeted and limited area of image modification (which a common natural assumption in image anonymization problems) and b/ additional constraint of image classifier invariance. In other words, we consider image classifier as a subjective measure of information preservation. From a human point of view, it is not a good measure that is easy to be fooled [9]. However, from the application point of view this a good measure that ensures that anonymization mapping is coherent with classification pattern model represented by image classifier.

In [20] Ren et al. train a model for a specialized task of face anonymization. They use generative adversarial training with a discriminator that tries to obtain private information from anonymized videos. Obtained anonymizer replaces a face with a similar, non-identifiable face, and thus it minimizes the perturbation effect on the object detector. Such approach is similar to ours in terms of building an anonymization algorithm which would affect the main task model (detection, in this case) as little as possible. However, the solution is different as we propose a general approach that does not require an anonymizer model training.

Leroux et al. proposed an obfuscation framework [12] that produces a non human-readable input that can be successfully treated by the task-specific neural network that was trained on non-obfuscated data. This framework consists of an obfuscator, de-obfuscator, and a fixed pretrained classifier. The obfuscator and de-obfuscator are trained in the adversarial manner.

In the similar spirit Li et al. proposed an adversarial training framework called DeepObfuscator [13]. It consists of an obfuscator, a classifier, an adversary reconstructor and an adversary classifier. When training, obfuscator aims to transform the image to defend from reconstruction attack and consequent adversary classification, while maintaining good performance on classifier. We remark that this approach also requires a model training, however the result is lightweight enough to produce smartphone-deployable obfuscator according to the paper.

In their work [21] Ryoo et al. introduced the concept of inverse super resolution (ISR). They developed a CNN model capable of classifying human activity at extreme low resolution videos, i.e. 16x12 pixels. Such input size allows to perform privacy-preserving classification.

2.2. Gradient-based image modification

As the focus of this work is the effect of input modifications on the model, it is also closely related to *adversarial attacks*, a research area that studies neural networks vulnerability to small input perturbations [9]. However it is worth stating that it does not fit exactly under this category because for obvious reasons the changes in the images introduced by anonymization procedure must be perceivable by human (indeed, if after the anonymization no change in the image is seen by the human eye, then from the human point of view the image contains same information as before). A recent survey on adversarial attacks is given by [5], while [30] reviews adversarial attacks in computer vision field. Adversarial attacks can be mostly categorized into two groups:

Black-box attacks: black-box attacks do not assume any knowledge about the model and uses information about inputs and outputs to conduct the attack [3, 14, 18];

White-box attacks: In a white-box attack, the adversary has total knowledge and access to the model being attacked [15, 16, 29].

Those types of attack can lead to privacy leaks. For instance, provided a white-box access to the network model Fredrikson et al. used model invertion to reconstruct a face image used in the training set using the confidence score of the target model [8].

2.3. Robust training

Learning how to protect models from input attacks, making them more robust, less sensitive to input change allows to design privacy-preserving methods which would not affect the model performance.

Adversarial defense can be seen as a particular of neural network robustness. Recent surveys [2, 17, 24] give overview on the adversarial attack and defenses.

Among all possible applications, the medical field is the one particularly concerned with privacy related issue. Indeed, the potential benefits of using multi-national data sets across multiple institutions could allow to significantly improve the accuracy of current models. However, regulation prevents the use of cross-medical information. This is why Kaissis et al. [11] proposed an end-to-end framework for training deep learning models of CNNs while preserving the privacy of the patients' data used.

Most, if not all, techniques detailed in this section have a common point. They all require to either develop a new model to increase robustness, or to retrain the original model to learn from the modified inputs. However it is not always possible or desirable to substitute an existing computer vision model. The approach presented in this paper can be seen as a white-box attack [4] but with the objective of maintaining the model's performance on the the anonymized images rather than decreasing it. We aim to demonstrate that the input image can be modified to remove the privacy-related information without changing the original model performance trained on the non-altered images.

3. Methodology

In this section we state more formally the problem setting, explain the proposed method and describe the datasets we conduct the experiments on.

3.1. Datasets

For the experiments we use two datasets with associated classification problems. The first one is the ImageNet [6], which is used here to provide reproducible results. More specifically, we used the test set of ImageNet. The second one is the proprietary dataset associated with the vehicle classification problem at the toll, similar to the one described in [23]. This labeled dataset comprises 401879 images of vehicles passing the toll, where labeling is done *w.r.t.* 5 vehicle classes according to the toll payment tariff grid in France [1]. Class samples are provided in Fig. 2. It is used here to illustrate the practical implications of the proposed algorithm.

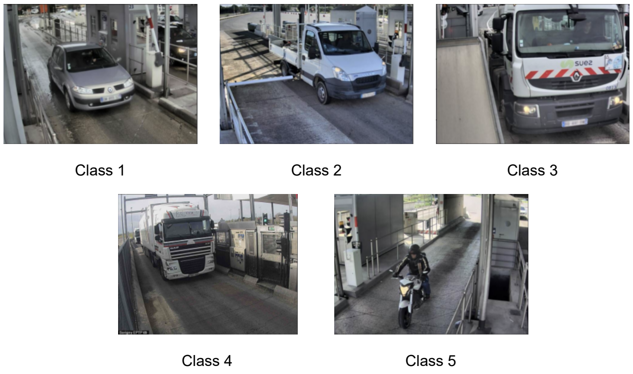
Class 1
Class 2
Class 3
Class 4
Class 5

Figure 2. Examples of vehicle classes in the proprietary dataset for toll vehicle classification task.

3.2. Anonymization

3.2.1 Anonymization algorithm

We use blur anonymization on a rectangular patch of the image containing sensitive information. The blur is done with Gaussian kernel of size 9 × 9 with $\sigma = 10$.

3.2.2 Anonymization area

For ImageNet dataset, we simulate the presence of sensitive data by selecting the 50×50 rectangular (assuming the image has been already resized) with high salience. In order to select such a rectangular we used a pretrained MobileNetV2 [22] to get the loss function gradient w.r.t. image and choose the rectangular with the highest sum of absolute gradient values. Some examples of resulting images are given in Fig. 3.

For proprietary vehicle dataset, we use an object detection model in order to identify the bounding boxes with the human faces and license plates present in the image. Their average width and height in the resized 224×224 images are 8 and 13 for the faces, and 20 and 11 for the license plates, respectively.

3.3. Problem formulation and method description

Let $x_{or} \in X$ be an original image, x_{blur} be an image obtained from x_{or} by blurring (or, more generally, modifying) an area corresponding to an indices subset I. Let $f : X \to [0, 1]^N$ be a neural network classifier that outputs the score vector associated with N classes classification task. The task is to find an image $\tilde{x}_{blur}$ such that

$$\mathrm{d}(f(\tilde{x}_{blur}), f(x_{or})) < \varepsilon, \qquad (1)$$

where distance d can be defined in multiple ways, notably as

- an indicator function that returns 1 iff $\tilde{x}_{blur}$ and x_{or} are mapped by f to the same class;
- euclidean distance.

We additionally require that image $\tilde{x}_{blur}$ is different from image x_{blur} only on indices subset I. It is implied (although we will not verify this explicitly) that $\tilde{x}_{blur}$ is in the vicinity of image x_{blur}, in particular, we expect $\tilde{x}_{blur}$ to remain far from x_{or} on I.

We remark that problem formulation (1) can be rewritten as an optimization problem *w.r.t* input image $\tilde{x}_{blur}$:

$$\tilde{x}^*_{blur} = \underset{\tilde{x}_{blur} \in X^I_{x_{or}}}{\arg\min} \; \mathrm{d}(f(\tilde{x}_{blur}), f(x_{or})) \qquad (2)$$

where $X^I_{x_{or}} = \{x \in X \,|\, x[i] = x_{or}[i] \,, i \notin I\}$. In this case, condition (1) plays a role of a stopping criterion. If we explicitly demand that $\tilde{x}_{blur} \in X^I_{x^*} \cap \mathbb{B}_\delta(x_{blur})$ in order to reflect the expectation for solution $\tilde{x}_{blur}$ to remain within small distance δ to starting point x_{blur}, the similarity to the adversarial attack problem [9, 26] becomes clear. The difference lies in the optimization goal: whereas an adversarial attack is searching for image x that maximizes distance $\mathrm{d}(f(x), f(x_{or}))$ for initial image x_{or}, adaptive anonymization algorithm minimizes distance $\mathrm{d}(f(\tilde{x}_{blur}), f(x_{or}))$, using x_{blur} as a starting point. To solve this problem we

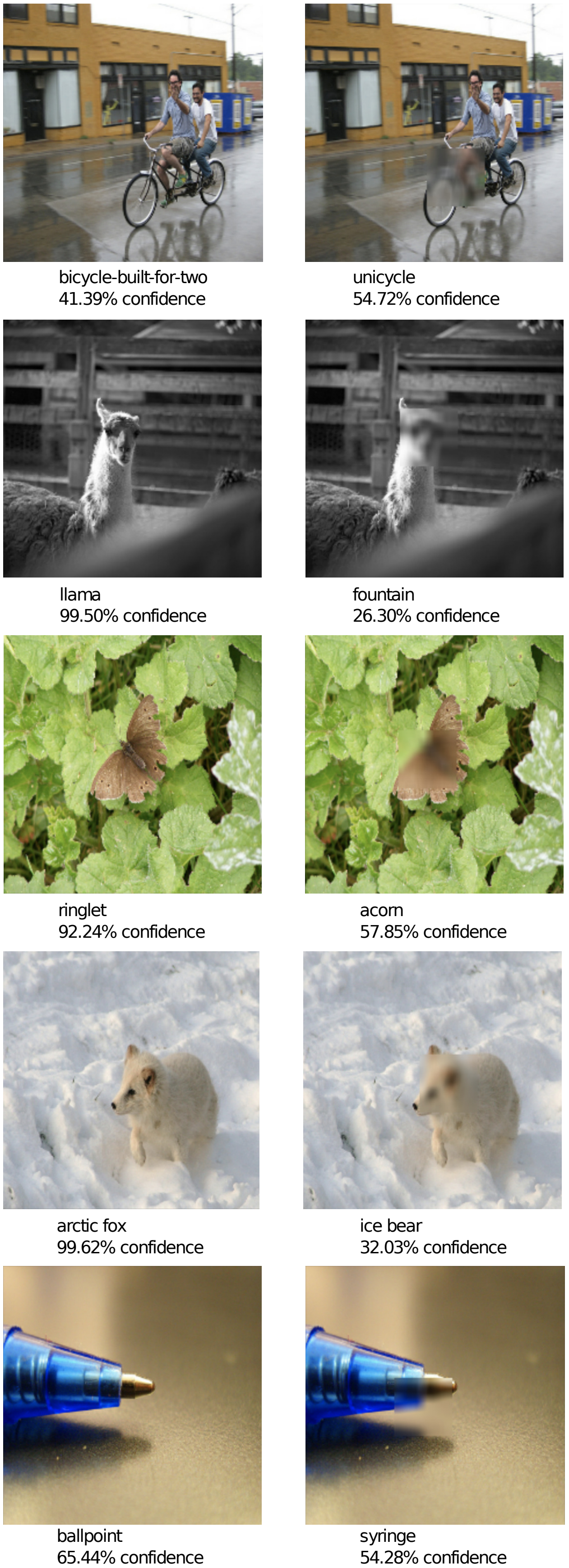

Figure 3. Examples of anonymizing simulated sensitive data areas in ImageNet test dataset. The predicted class and score corresponds to the pretrained MobileNetV2 inference results.

propose to use the gradient descent method by minimizing the cross-entropy of $f(\tilde{x}_{blur})$ and $f(x_{or})$ with respect to $\tilde{x}_{blur}[i]$, $i \in I$. More specifically, we construct $\tilde{x}_{blur}$ in an iterative way as:

$$\begin{aligned} \tilde{x}_{blur}^{(0)} &= x_{blur} \\ \tilde{x}_{blur}^{(k)} &= \tilde{x}_{blur}^{(k-1)} - \alpha \nabla_I J(\theta, \tilde{x}_{blur}^{(k-1)}, f(x_{or})), \end{aligned} \tag{3}$$

where J denotes cross-entropy, θ is the set of model parameters, and gradient is taken with respect to the I entries of image $\tilde{x}_{blur}^{(k-1)}$. The iterative process continues until the stopping criterion 1 is met, or the maximum number of iterations are done. Diagram representation of the proposed method in given in Fig. 4.

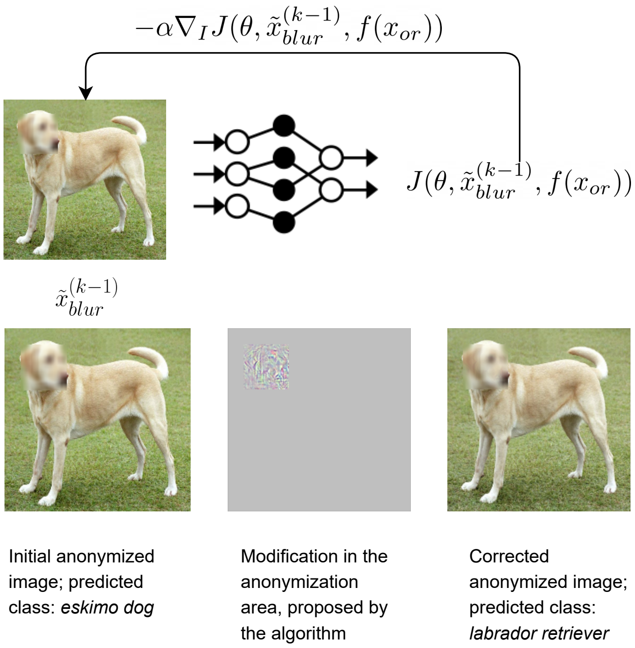
$-\alpha \nabla_I J(\theta, \tilde{x}_{blur}^{(k-1)}, f(x_{or}))$
$J(\theta, \tilde{x}_{blur}^{(k-1)}, f(x_{or}))$
$\tilde{x}_{blur}^{(k-1)}$
Initial anonymized image; predicted class: *eskimo dog*
Modification in the anonymization area, proposed by the algorithm
Corrected anonymized image; predicted class: *labrador retriever*

Figure 4. Proposed method iteratively "corrects" anonymized version of the image aiming to obtain the same model prediction as on the original image. Image change is obtained via gradient descent for loss function w.r.t. area of the image subject to modifications.

In our experiments we have used a predicted class equality indicator as the distance function and we set maximum number of iteration to 100. We normalize the values of the gradient (by diving the gradient by its entry of the maximum absolute value) in order to consistently reflect the step size by parameter α.

4. Results and discussions

In this section we test the efficiency of the proposed method on two different datasets and associated classification tasks.

4.1. Experiments on ImageNet

We start by the sensitivity analysis of the proposed method to the step size α for the given CNN MobileNetV2. Next, we validate method performance for multiple CNN architectures.

4.1.1 Sensitivity to the step size α

In order to study the method sensitivity to the step size α and to estimate empirically its ”good” value, we have conducted a series of experiments varying only this parameter. More precisely, we considered MobileNetV2 network pretrained on ImageNet dataset and a range of values for the parameter α from 0.025 to 2. The results of these experiments are presented in Table 1. Initially, due to purposely severe anonymization simulation, a large part of test set turned out to be vulnerable to this type of ”attack”: 15.944% of images changed the predicted class after applying anonymization. Then, we applied the proposed method with various values of the parameter α. We observed that for all experiments only of a small part of vulnerable images resist to the correction (from 0% to 0.301%). Therefore, the percentage of vulnerable images decreases significantly from 15.944% to 0%-0.048% (depending on the α value). The best results were achieved with the values of $\alpha = 0.025, 0.05, 0.1, 0.75$ where no images remained vulnerable after the application of the algorithm. Overall, one can conclude that the proposed algorithm has low sensitivity to parameter α and shows good performance for the whole range of tested values.

4.1.2 Experiments on various CNN architectures

After observing only moderate influence of the parameter α on the method performance we can proceed to the a study on various CNN architectures without worrying much about precisely estimating the best parameter settings. We have considered the following architectures: MobileNetV2, ResNet50V2, InceptionV3, Xception. As before, we use the neural networks pretrained on ImageNet dataset, and compare their predictions on the original and modified (augmented with the blur) test set of ImageNet. The step size α is fixed at 0.1. The results of this experiment are given in Table 2. We observe that

- all of the tested architectures showed vulnerability to this sort of “anonymization attack”. The share of vulnerable images varied from 9.86% for Xception to 16.44% for ResNetV2.

- regardless of the architecture the proposed method had performed well and has fully accomplished the set goal. In all four experiments the share of vulnerable images was reduced to 0%.

- there is a tangible gap between initial share of vulnerable images for MobileNetV2 and ResNetV2 (15.94% and 16.44% respectively) vs InceptionV3 and Xception (11.86% and 9.86%). Similar dependency is observed in the variation of the experiment with masking applied instead of blurring (see subsection 4.1.3). We suggest that this is related to CNN architecture type, but further research is needed to explore this hypothesis. Thus we leave it as an observation for a curious reader.

- the maximum number of iterations for the algorithm was set to 100. However in reality the number of steps required to go back to the initially predicted class was much lower. In particular, for MobileNetV2 average number of steps was equal to 1.95, 1.93 steps for ResNet50V2, 2.00 steps for InceptionV3, and 1.81 steps for Xception (here the average is taken only over images that needed correction).

Conducted experiments validate the proposed method and demonstrate its applicability for a range of different initial settings.

4.1.3 Using masking as an extreme anonymization technique

Taking into account good performance of the proposed method for Gaussian blur, we have decided to go further with more severe anonymization method, namely masking. In this set of experiments we keep the same settings as in the experiments with different CNN architectures (see subsection 4.1.2), but instead of applying Gaussian blur on the desired area we replace it with an independently generated patch, in particular we assign single value to all the pixels in that region. In this scenario we are confident that the information in the desired region has been removed completely, and no partial remaining information is able to indirectly aid the proposed method in achieving the task. The results of such experiment is given in Table 3. We observe the increased number of vulnerable images which is aligned with the intuition of masking being a more radical type of anonymization than blurring. Nevertheless we observe a good algorithm performance with the vast majority of anonymized images modified to correspond to the initially predicted class. In particular, for MobileNetV2 1 image of 23207 initially vulnerable ones remained uncorrected, for ResNet50V2 7 images out of 26433, for Inception V3 1 image out of 16027, and 0 images out of 14072 for Xception. However, an increased number of algorithm iteration steps was required to reach the stopping criterion in comparison with the previous experiment with the blurring used for anonymization. On average, 3.54 iteration steps were required for MobileNetV2 (+ 84%), 4.69 steps

Table 1. Proposed method performance w.r.t. varying step size α. Experiments were conducted with MobileNetV2 on the test set of ImageNet consisting of 100k images. Initially 15,944 images changed predicted class after anonymization, which constitutes 15.94% of the test set.

step size	# images that changed predicted class after anonymization	# images not corrected by the algorithm	% images not corrected by the algorithm	% vulnerable images (initial)	% vulnerable images (final)
0.025	**15944**	**0**	**0.000%**	**15.94%**	**0.000%**
0.05		**0**	**0.000%**		**0.000%**
0.1		**0**	**0.000%**		**0.000%**
0.25		1	0.006%		0.001%
0.5		1	0.006%		0.001%
0.75		**0**	**0.000%**		**0.000%**
1		5	0.031%		0.005%
1.5		9	0.056%		0.009%
2		48	0.301%		0.048%

Table 2. Proposed algorithm performance w.r.t. different CNN architectures. Experiments were conducted on the test set of ImageNet consisting of 100k images. For all the architectures considered number of images vulnerable to the predicted class changed was reduced to 0.

Architecture	# images that changed predicted class after anonymization	% vulnerable images (initial)	% vulnerable images (final)	average number of the iterations
MobileNetV2	15944	15.94%	0.000%	1.95
ResNet50V2	16436	16.44%		1.93
InceptionV3	11863	11.86%		2.00
Xception	9864	9.86%		1.80

for ResNet50V2 (+ 143%), 3.54 steps for InceptionV3 (+ 77%), and 3.38 steps for Xception (+ 88%).

4.1.4 Adaptive face anonymization for ImageNet dataset

A recent publication [28] provides face annotations for ImageNet dataset. We used them with the anonymization function given in [28] to conduct experiments similar to Section 4.1.2. The results obtained (Table 4) are well aligned with the results presented in Table 2 both in method efficiency (0% failure) and average number of iterations and steps (1.91-2.20). Additionally, we remark that a significant amount of ImageNet test images containing faces (17.5 %), and a noticeable amount of images vulnerable to class switching after initial anonymization (2.0%-3.3%) indicate suitability of this dataset for the considered problem.

4.1.5 Experiment with a Transformer architecture

Motivated by the fact that Transformer [27] architectures currently achieve SOTA in various computer vision tasks including image classification, we have conducted an experiment similar to those in the Section 4.1.2 with a pre-trained ViT-B/16 model [7]. Out of the 100k ImageNet test images, 5901 images have changed the predicted class after the anonymization with the Gaussian blur. Proposed method was able to successfully undo the change for all the affected images with the average of 2.41 iteration steps required. We observed that this result is consistent with the experiments results for MobileNetV2, ResNet50V2, InceptionV3 and Xception provided in Table 2.

4.2. Experiments on real-world application of vehicle classification at the toll

For this experiment, we use a customized model from VGG family [25] trained on the dataset with presumably the same data distribution as the test set. On the test dataset the model has reached 93.30% accuracy, whereas on the blurred images accuracy has decreased to 93.24%. In particular, on 725 images the class predicted by the model has changed. This amount of images comprises 0.18% of the whole dataset, much less than in the experiment on Im-

Table 3. Proposed algorithm performance w.r.t. different CNN architectures, where anonymization has been done with masking.

Architecture	# images that changed predicted class after anonymization	# images not corrected by the algorithm	% vulnerable images (initial)	% vulnerable images (final)	average # iterations
MobileNetV2	23207	1	23.21%	0.001%	3.58
ResNet50V2	26433	7	26.43%	0.007%	4.69
InceptionV3	16027	1	16.03%	0.001%	3.54
Xception	14072	0	14.07%	0.000%	3.38

Table 4. Proposed algorithm performance for adaptive face anonymization task on ImageNet

Architecture	% vulnerable images (initial)	% vulnerable images (final)	average # iterations
MobileNetV2	3.14%	0.00%	2.07
ResNet50V2	3.31%	0.00%	2.20
InceptionV3	2.34%	0.00%	2.00
Xception	2.01%	0.00%	1.91

ageNet. Indeed, in our ImageNet experiment in order to test the limits of the proposed method we are simulating an ”aggressive” anonymization which on purpose targets the sensitive image areas and which perturbs rather large parts of the image. Additionally, the vehicle classification problem has much lower number of classes (5 vs. 1000 for ImageNet), which further decreases the probability of model output class change. Here we see a real life example where the problem is not as accentuated. Nevertheless, under the conditions of large data volumes (which is exactly the case with the vehicle classification at the toll) even a small accuracy decrease can lead to tangible money and reputation loss for the client. Hence, the accuracy control on the anonymized data requires ideally the full consistency of the classification model on the original and anonymized datasets.

Initial accuracy on the abovementioned vulnerable subdataset was 64.55%, whereas after the blurring it has dropped dramatically to 31.59%. After replacing the Gaussian blur with the proposed method we observe 100% of classification inconsistency correction. In other words, the classification model has regained back its initial accuracy of 93.30%. Out of 725 images that needed adaptive anonymization, 680 images were ”corrected” in 1 iteration step, 41 images required 2 iteration steps, and 4 images needed 3 iteration steps. We observe that despite setting a high value of maximal number of iteration steps (100), in the actual experiment a low number of steps were required to perform the anonymization correction.

4.3. Discussions

Throughout the method development and experiment conducting we have made a couple of observations we would like to share here.

- Presented method at its core has a simple gradient descent. Consequently, it can be improved (where the notion of improvement is tied to the specific problem tasks and limitations) by substitution of gradient descent by any of more efficient gradient methods. In the present work we did not set as a goal to find the most performing variation of the method, but rather to present the concept as a whole.
- The stopping criterion could be easily substituted by precision of approximation to the original image class score, or precision of approximation to the original image probability vector.
- From the theoretical point of view nothing prevents application of the proposed method to other deep learning tasks.

5. Conclusions

In this work we have considered the problem of anonymizer-classifier pipeline robustness for the task of image classification from the non-typical viewpoint of modifying the anonymizer rather than classification model. Besides general advantage of such direction which lies in initial classification model preservation, proposed approach has a number of benefits, such a implementation simplicity, no dependency on specific neural network architecture, low number of parameters, and, consequently, the possibility to use it as a “plug-and-play” method. As a disadvantage, increased number of resources are required during the method usage as it needs multiple inference runs of the classifier model. Conducted experiments show good performance of the proposed method and low sensitivity to its parameter, the step size.

References

[1] Vehicle Classification. https://www.autoroutes.fr/en/vehicle-classification.htm. [Online; accessed 19-June-2021]. 4

[2] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access*, 6:14410–14430, 2018. 3

[3] Arjun Nitin Bhagoji, Warren He, Bo Li, and Dawn Song. Practical black-box attacks on deep neural networks using efficient query mechanisms. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 158–174, Cham, 2018. Springer International Publishing. 3

[4] Anirban Chakraborty, Manaar Alam, Vishal Dey, A. Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey. *ArXiv*, abs/1810.00069, 2018. 4

[5] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. A survey on adversarial attacks and defences. *CAAI Transactions on Intelligence Technology*, 6(1):25–45, 2021. 3

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 7

[8] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, page 1322–1333, New York, NY, USA, 2015. Association for Computing Machinery. 3

[9] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 3, 4

[10] Abul Hasnat, Nadiya Shvai, and Amir Nakib. Cnn classifier's robustness enhancement when preserving privacy. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 3887–3891. IEEE, 2021. 1, 2

[11] Georgios Kaissis, Alexander Ziller, Jonathan Passerat-Palmbach, Théo Ryffel, Dmitrii Usynin, Andrew Trask, Ionésio Lima, Jason Mancuso, Friederike Jungmann, Marc-Matthias Steinborn, Andreas Saleh, Marcus Makowski, Daniel Rueckert, and Rickmer Braren. End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nature Machine Intelligence*, 3:1–12, 06 2021. 3

[12] Sam Leroux, Tim Verbelen, Pieter Simoens, and Bart Dhoedt. Privacy aware offloading of deep neural networks. 2018. 3

[13] Ang Li, Jiayi Guo, Huanrui Yang, and Yiran Chen. Deepobfuscator: Adversarial training framework for privacy-preserving image classification. *ArXiv*, abs/1909.04126, 2019. 3

[14] Huiying Li, Shawn Shan, Emily Wenger, Jiayun Zhang, Haitao Zheng, and Ben Zhao. Blacklight: Defending black-box adversarial attacks on deep neural networks, 06 2020. 3

[15] H. Liu, Zhenyu Zhou, Fanhua Shang, Xiaoyu Qi, Yuan yuan Liu, and L. Jiao. Boosting gradient for white-box adversarial attacks. *ArXiv*, abs/2010.10712, 2020. 3

[16] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 739–753, 2019. 3

[17] Mesut Ozdag. Adversarial attacks and defenses against deep neural networks: a survey. *Procedia Computer Science*, 140:152–161, 2018. 3

[18] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, ASIA CCS '17, page 506–519, New York, NY, USA, 2017. Association for Computing Machinery. 3

[19] Protection Regulation. Regulation (eu) 2016/679 of the european parliament and of the council. *REGULATION (EU)*, 679:2016, 2016. 1

[20] Zhongzheng Ren, Yong Jae Lee, and Michael S. Ryoo. Learning to anonymize faces for privacy preserving action detection. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 639–655, Cham, 2018. Springer International Publishing. 3

[21] Michael S. Ryoo, Brandon Rothrock, Charles Fleming, and Hyun Jong Yang. Privacy-preserving human activity recognition from extreme low resolution. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4255–4262. AAAI Press, 2017. 3

[22] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 4

[23] Nadiya Shvai, Abul Hasnat, Antoine Meicler, and Amir Nakib. Accurate classification for automatic vehicle-type recognition based on ensemble classifiers. *IEEE Transactions on Intelligent Transportation Systems*, 21(3):1288–1297, 2019. 4

[24] Samuel Henrique Silva and Peyman Najafirad. Opportunities and challenges in deep learning adversarial robustness: A survey. *arXiv preprint arXiv:2007.00753*, 2020. 3

[25] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 7

[26] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 4

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 7

[28] 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 (ICML)*, 2022. 7

[29] Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, 30(9):2805–2824, 2019. 3

[30] Yiyun Zhou, Meng Han, Liyuan Liu, Jing He, and Xi Gao. The adversarial attacks threats on computer vision: A survey. In *2019 IEEE 16th International Conference on Mobile Ad Hoc and Sensor Systems Workshops (MASSW)*, pages 25–30, 2019. 3