sensors

MDPI

Article

Lie to Me: Shield Your Emotions from Prying Software

Alina Elena Baia [1], Giulio Biondi [2], Valentina Franzoni [2,3,*], Alfredo Milani [2] and Valentina Poggioni [2,*]

[1] Department of Mathematics and Computer Science, University of Florence, Viale Morgagni 67/a, 50134 Florence, Italy; alinaelena.baia@unifi.it
[2] Department of Mathematics and Computer Science, University of Perugia, Via Vanvitelli 1, 06123 Perugia, Italy; giulio.biondi@unipg.it (G.B.); alfredo.milani@unipg.it or milani@unipg.it (A.M.)
[3] Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong, China
* Correspondence: valentina.franzoni@dmi.unipg.it (V.F.); valentina.poggioni@unipg.it (V.P.)

Abstract: Deep learning approaches for facial Emotion Recognition (ER) obtain high accuracy on basic models, e.g., Ekman's models, in the specific domain of facial emotional expressions. Thus, facial tracking of users' emotions could be easily used against the right to privacy or for manipulative purposes. As recent studies have shown that deep learning models are susceptible to adversarial examples (images intentionally modified to fool a machine learning classifier) we propose to use them to preserve users' privacy against ER. In this paper, we present a technique for generating Emotion Adversarial Attacks (EAAs). EAAs are performed applying well-known image filters inspired from Instagram, and a multi-objective evolutionary algorithm is used to determine the per-image best filters attacking combination. Experimental results on the well-known AffectNet dataset of facial expressions show that our approach successfully attacks emotion classifiers to protect user privacy. On the other hand, the quality of the images from the human perception point of view is maintained. Several experiments with different sequences of filters are run and show that the Attack Success Rate is very high, above 90% for every test.

Keywords: emotion recognition; adversarial machine learning; privacy protection; evolutionary algorithm

Citation: Baia, A.E.; Biondi, G.; Franzoni, V.; Milani, A.; Poggioni, V. Lie to Me: Shield Your Emotions from Prying Software. *Sensors* **2022**, *22*, 967. https://doi.org/10.3390/s22030967

Academic Editor: Raffaele Gravina

Received: 1 January 2022
Accepted: 22 January 2022
Published: 26 January 2022

1. Introduction

Visual Emotion Recognition (ER) is one of the first Affective Computing techniques [1] that have been widely studied in computer science and artificial intelligence, based on visual features of the facial expression. Several different approaches at visual recognition obtained different grades of classifications, from face landmarks to ER using deep learning and knowledge transfer [2–5], which is currently the most common approach to facial ER. Using a Convolutional Neural Network (CNN) to detect basic emotions from an image or video frame, the resulting accuracy in the Ekman model of emotions is promising and has been implemented already [6]. Facial ER [2,6] can obtain excellent results with relatively small data sets of images, when trained on a single individual. Many are the ethical application areas in research of facial ER, e.g., to detect particular states needing an immediate medical intervention, or changes over time underlying a degenerative health condition. On the other hand, the most common applications of facial ER are prone to potentially unethical manipulation of users' preferences. In behavior-tracking applications, the emotional reactions of a user in front of a product could produce extremely precious insights for companies, government, or political parties, prying into the user's habits and emotional states. e.g., marketing applications in a supermarket, in front of a shop showcase, or browsing an e-commerce website [7,8]; tracking drivers' states [9]; analyzing pieces of information in social networks [10]; analyzing news or political opinions [11]; military robot interaction [12]). For the critical nature of such information, tracking it could open a breach in personal data confidentiality, and become a potential source of manipulation bias for the user's preferences.

The diffusion and the wide use of deep learning methods for artificial intelligence systems, thus, pose significant security and privacy issues. From the security point of view, Adversarial Attacks (AA) showed that deep learning models can be easily fooled [13–25] while, from a privacy point of view, it has been shown that information can be easily extracted from dataset and learned model [26–28]. It has also been shown that attacking methods based on adversarial samples can be used for privacy-preserving purposes [29–33]: in this case, data are intentionally modified to avoid unauthorized information extraction by fooling the unauthorized software.

With the increasing popularity of online social networks, privacy-preserving photo sharing has received considerable attention, and several systems have been proposed [33–36].

In this paper, we propose a technique for Emotion Adversarial Attack (EAA) to filter out the emotional features from video frames or photos of human faces to ensure the users' freedom and protection against emotion recognizers in environments where they may be unauthorized and therefore prying.

The fooling protecting filters are built by composing and parametrizing popular image-enhancing Instagram filters: they are the result of an optimization process implemented by a nested-evolutionary algorithm [13,14,33]. Applying these protecting filters to any image, we obtain a series of other images from which information extraction is more difficult.

Since the algorithm works in a black-box scenario, it does not require any information about the model's parameters or gradient values, as many other systems require.

The proposed algorithm combines the idea of a Multi-Objective Evolutionary (MOE) approach for adversarial attacks [13] with the *per-instance* approach presented in [33]. Compared to the MOE approach, we also introduce the use of a Full-Reference Image Quality Assessment (FR-IQA).

The *per-image* approach allows the discovery of personalized sequences of filters having different image-specific characteristics; the image assessment allows creating high quality and natural-looking adversarial privacy-preserving samples. The preferred aspect can be chosen by the user, while the image quality is controlled by the multi-objective fitness function implemented by the Structure Similarity Index (SSIM) [37].

This approach allows to overcome the main flaws of restricted attack methods that in general produce not semantically meaningful modifications that are easily detectable by software, even if they are imperceptible by human eyes [38–40].

Moreover, performing the attack using well-known filters widely used in social media (e.g., Instagram) makes our filter composition indistinguishable from any other filter composition extensively used every day to enhance photos and images. This approach essentially makes our attacks transparent to the human perception, still keeping their privacy-preserving emotional features.

We tested the algorithm on the AffectNet data set [41] varying the length of the sequence (3, 4 and 5 filters) obtaining attack success rates up to 96%.

2. Background

2.1. Emotion Adversarial Attacks

For an input facial image $x \in X \subset \mathbb{R}^d$ and the related label y, let F be a Neural Network (NN) classifier that correctly predicts the emotional class label for the input image $x : F(x) = y$. An EAA attempts to modify x adding a δ perturbation into an adversarial image $x^* = x + \delta$, such as to induce F to make a *faulty emotion class prediction*, i.e., $F(x^*) \neq F(x)$.

If we consider the type of perturbation applied δ, the attacks can be classifiable as either *restricted* or *unrestricted*. If restricted, the changes applied to the original image are typically small and bounded by a L_p-norm distance, forcing the adversarial image x^* to be as similar as possible to the initial input. On the other hand, unrestricted attacks use large unconstrained L_p-bounded perturbations manipulating the image to create adversary photorealistic instances. In this case, the intent is not to restrict the transformations on pixels but to limit the human perception that a transformation has been applied [17,42,43].

2.2. Image Filters

Inspired by Instagram, which offers tools to seamlessly modify images, we propose to combine multiple image filters to create custom adversarial image transformations. This approach provides plenty of styles options, ranging from subtle and warm looks to more dramatic and vivid colors effects.

As proposed in [13,14,33], we have used popular Instagram image filters such as *Clarendon, Juno, Reyes, Gingham, Lark, Hudson, Slumber, Stinson, Rise,* and *Perpetua*. Each filter has distinct properties and aspects, such as different contrast, saturation, brightness, and shadow levels. These differences allow the production of different effects that are usually composed by the users, e.g., *Rise* mixes a radial gradient with a sepia hue, while *Clarendon* brightens and highlights the image, *Juno* increases saturation, and Gingham provides a vintage appearance.

Each filter is parameterized by two values to be optimized by the algorithm: *intensity* α and *strength* s. The role of α is to alter the intensity of each filter component, e.g., contrast, saturation, brightness, gamma correction, edge enhancement.

The s parameter is used to manage the filter impact, defined as the convex interpolation between the input photo x and the altered image x^*:

$$strength(x, x^*, s) = (1.0 - s) \cdot x + s \cdot x^* \tag{1}$$

The cases are $s = 0$, where the image is not altered by the filter, and $s = 1$ where the filter returns a mutated image x^*.

2.3. Image Quality Assessment

Image quality assessment (IQA) techniques are used to quantify the visual quality of an image by analyzing different characteristics such as aesthetics, naturalness, or distortions [44–46]. IQA methods are used for a variety of applications, ranging from benchmarking image processing algorithms or monitoring image quality to optimizing algorithms in the context of visual communication systems. Over the years, many different methods have been proposed. There are essentially two types of IQA methods, *subjective* and *objective*. Subjective assessment requires a human evaluation and intervention and is considered the most accurate and reliable. However, it is time-consuming, expensive, and impractical for real-time assessment applications.

Objective methods are designed to measure the visual quality of an image automatically fitting the human assessment. Using mathematical models or deep learning approaches, they result highly efficient and ideal for image-based system optimization.

Based on the availability of the reference image, objective methods can be further divided into three categories: *Full-Reference* (FR), *Reduced-Reference* (RF), and *No-Reference* (NR). FR strategies require computing the quality score by comparing the modified image with the complete reference image. RF strategies use only partial information from the reference image, such as extracted features. NR strategies, also known as blind assessments, are designed to accurately predict the image quality without using a reference image or any additional information, thus being suitable for applications where the reference image is not available.

Given the configuration of the proposed algorithm and the availability of clean reference images in our work, we use the SSIM index [37], a well-known and well-performing FR-IQA method to automatically and objectively assess the quality of adversarial samples generated by the EAA.

SSIM, introduced by Wang et al. [37], is an FR perceptual metric that quantifies the image degradations as perceived changes in the structural information. SSIM is a content-aware assessment metric, inspired by the Human Visual System (HVS), capable of extracting and identifying structural information from natural scenes (i.e., images), deeply structured with significant dependencies between spatially closed pixels. A measure that exploits the characteristics of the HVS can better match the subjectively-perceived visual

quality. Capturing the change in structural information provides a good approximation of the perceived image degradation.

Structural information is defined as attributes describing objects independent from luminance and contrast. In other words, SSIM is a structural similarity measure that compares patterns of pixels intensities normalized for contrast and brightness, which variability does not alter the structures of the objects in the images. Given that different regions of an image may have different levels of contrast and luminance, the SSIM index is computed locally within a predefined 1-pixel local window, and the overall image quality is evaluated by taking the mean for the number of local windows of the image. The SSIM index is thus a multiplicative combination of three terms of comparison: luminance, contrast, and structure, computed over the image's patches. SSIM was designed to satisfy symmetry, boundedness (i.e., where the score is bounded by an upper value equal to 1), and unique maximum property where the SSIM score is equal to 1 if and only if the two compared images are identical. In general, with a score value higher than 0.99, the images are considered to be indistinguishable (Algorithm 1).

Algorithm 1: AGV-emotion-attack

Input: Image I, population size N, generations E
Initialize population P of N individuals;
Evaluate each individual of P by *EASR* and *SSIM*;
for $e = 0$ **to** E **do**
  Offsprings = $\{\emptyset\}$;
  for $i = 1$ **to** N **do**
    Select randomly $parent_1$,$parent_2$ from P ;
    $\overline{p}_1 \leftarrow encode_1(parent_1)$;
    $\overline{p}_2 \leftarrow encode_1(parent_2)$;
    y_i = crossover($\overline{p}_1$,$\overline{p}_2$);
    $\overline{y_i}$ = mutation(y_i) ;
    $n_i \leftarrow encode_2(\overline{y}_i)$;
    $\overline{n}_i$ = optimizer$_O(n_i)$;
    Offsprings $\leftarrow (\overline{y}_i, \overline{n}_i)$;
  end
  foreach $(\overline{y}_i, \overline{n}_i) \in$ Offsprings **do**
    $b \leftarrow$ decode $(\overline{y}_i, \overline{n}_i)$;
    Evaluate the fitness by means of *EASR* and *SSIM* ;
  end
  P = selection(P, Offsprings) ;
end
return: best sequence of filters for image I;

3. Related Works

Adversarial attacks to emotion recognition is a very recent application and just very few works are available in the literature [47–49]. The main difference with our work relies on the approach: white-box versus black-box. Since our algorithm works in a black-box scenario, it does not require any information about the model's parameters or gradient values, as the other systems require. Hence, our approach can be applied against any system without having any knowledge about it. Moreover, they also differ in the way the images are modified.

In particular [49] belongs in the category of physical attacks since it realizes attacks to facial biometric systems by printing a pair of eyeglass frames.

In [47,48], a saliency map extractor is used to extract the essential expression features of the clean facial expression example and a face detector is employed to find the position of the face in the image. This information is then used to enhance and cut the gradient of the input samples computed by the optimized momentum iterative method (OMIM) with respect to the misclassification loss.

4. Emotion Recognition Settings

An emotion recognizer, based on MobileNetV2 [50] with transfer learning, has been designed as the target of the emotion adversarial attack image generation algorithm, to deceive the emotion classifier, while maintaining a realistic human perception quality. The emotion recognizer, a CNN, classifies a human face in the seven basic emotions of the Ekman model of seven emotions [51], *Anger*, *contempt*, *disgust*, *fear*, *happiness*, *sadness*, *surprise*, extended with an eighth *neutral* class.

4.1. Data Set and Data Preparation

The AffectNet [41] data set, composed of 291,651 images labeled within the eight categories of the extended Ekman model, has been used to fine-tune the classifier (see Section 4.1). AffectNet is among the most widely used data sets for ER, and provides a large amount of images to be used for ER and selected for EAA. Since the categorical distribution in AffectNet samples is unbalanced, with *happiness* and *neutral* accounting together for about 2/3 of the data set, data have been randomly sampled to optimize classes with 3500 images per emotion, for a total of 28,000 images. An 80–20% proportion has been used in randomly splitting images in each category between the training and test sets. Data augmentation has been performed to optimize the training phase, applying random horizontal flipping and horizontal/vertical shifting to the images, by a random offset in the $[-15,+15]$ pixels range.

4.2. Emotion Recognizer Structure and Training

As with Transfer Learning (TL) the number of samples for NN training can be smaller than training a neural network from scratch, TL is particularly suitable in our case. TL typically allows adapting a NN (in our case, our convolutional neural network), pre-trained on a large image data set on several classes, to a network able to classify into a smaller set of possibly different categories, by using a smaller image data set and a faster training phase.

TL is based on using the pre-trained network structure and weights and replacing the last Fully-Connected (FC) classification layers with new, domain-specific layers and thus learning and adapting, i.e., *fine-tuning*, only their new weights. The idea behind this method is that different layers in a deep convolutional neural Network take into account different features; in particular, the top layers consider domain-agnostic primitive visual features, e.g., lines, and pixel color features. Deeper layers recognize more complex shapes and color distributions; the last layers in the network are responsible for assembling, by learning appropriate weights, the previously learned features into domain-specific information used for general image classification. Replacing the final layers allows the network to keep its low-level features recognition ability, saving training time, and adapting it to a new domain by re-training the new layers only.

In order to build the emotion recognizer by transfer learning, the MobileNetV2 [50] network was chosen, for its relatively small number of parameters, i.e., size. MobileNetV2 is pre-trained on the ImageNet data set [52,53] and it is able to classify images into 1000 categories. It is relevant to notice that different neural networks can have different performance on different data sets, and chosing a commonly-used data set as AffectNet, and a NN with a low number of parameters enhances the experiment clarity, still allowing future comparison with other approaches, datasets and networks. In the emotion recognizer deep network structure, the last MobileNetV2 fully connected classification layer is replaced with a FC layer of size 128 followed by a 0.5 dropout layer and a final fully connected layer of size 8 for the emotion classes, where the FC layers activation functions are, respectively, ReLU and Softmax.

Cross-validation in the CNN training is used to find the best set of hyperparameters obtained by the optimizer for given data. We specified the mini-batch size to 10, and validation data are shuffled at the beginning of each epoch. An epoch is a full training cycle on the entire training data set (i.e., 80% for our hold-out split).

During the training of the ER, starting with the initial pre-trained MobileNetV2 weights, we used the Stochastic Gradient Descent with Momentum (SGDM) as an optimizer, a *piecewise decay* optimizer policy, and a variable learning rate from an initial value of 1×10^{-3}, halving every 10 epochs on a total of 80.

5. Algorithm for Adversarial Attacks

The algorithm used to produce the attacks is implemented by mixing the idea of the MOE approach proposed in [13] and the *per-instance* approach proposed in [33]. We decided to use the MOE approach, which allows maximizing the method effectiveness and minimizing the image distortion. With a simpler optimization criterion, e.g., using the attack success rate only, the images could be excessively modified, creating an unnatural look. The description of the algorithm is given in Algorithm 1.

The optimization method consists of two nested evolutionary algorithms: an *outer algorithm*, using a generative adversarial approach based on a genetic algorithm, in charge of finding the sequence of filters to use, and an *inner algorithm*, based on Evolution Strategy (ES), used to choose the values of parameters. Given a set $S = \{f_1, f_2, \cdots f_m\}$ of m image filters, the outer algorithm genotype (with length l) is encoded as a list of filters, while the inner algorithm genotype is represented by a list containing the parameters for each selected filter.

5.1. Outer Algorithm

The outer-algorithm optimization is performed by a genetic algorithm: a population of N candidates is iteratively evolved. The candidates are randomly chosen to breed a new generation by the crossover and mutation procedures, where the candidates are evaluated on their fitness. At the end of each iteration, the best candidates are selected for the next generation:

Initial population: Generated by randomly selecting l filters from the S available set, and their parameters are initialized to 1.
Crossover: We use a one-point crossover to generate new off-springs (i.e., children) from random members. Each child is assured of inheriting genetic information from both parents.
Mutation: A filter is replaced with another, on a probability of mutation. The new filter is initialized with random parameters, assuring their complete mutation.
Selection: At each iteration, the N best individuals are chosen from the set of 2N candidates (i.e., parents and offsprings), according to their fitness. The same process is repeated until the algorithm spends the fixed amount of generations. The selection is implemented as a multi-objective evolutionary problem based on two criteria: Attack Success Rate (ASR) and image quality (evaluated by SSIM). The addition of the image quality assessment in the population evaluation phase gives the algorithm the capabilities to create high-quality and natural-looking adversarial examples. Given F a target *facial emotion recognizer*, x_i an original facial image, x_i^* derived from x_i by applying a sequence of filters, the fitness function is evaluated by the following:

$$\mathcal{F}(x_i, x_i^*) = \{1.0 - EASR_i(x_i, x_i^*), 1 - SSIM_i(x_i, x_i^*)\}, \tag{2}$$

where $EASR_i$ is the emotion adversarial attack success Rate obtained by classifying the modified image x_i^* with the target emotion classifier:

$$EASR_i = \begin{cases} 1, \ if \ F(x_i) \neq F(x_i^*) \\ 0, \ otherwise, \end{cases} \tag{3}$$

and $SSIM_i$ represents the image quality score that controls the amount of the applied perturbation described in Section 2.3.

5.2. Inner Algorithm

The inner algorithm is committed to the optimization of parameters, accomplished by $(1, \lambda)$ evolution strategy with $\lambda = 5$. A search distribution is iteratively updated by ES, following the gradient towards increased expected fitness. For each list of parameters, we compute N candidates through a perturbation of the original individuals. The gradient is estimated to better solutions comparing the fitness values of the N candidates. The gradient is then used to replace the previous individual. The entire process is repeated until meeting a stopping criterion.

6. Experiments and Discussion

6.1. Experimental Setup

The proposed algorithm has been evaluated by attacking the *Emotion Recognizer* described in Section 4, which is based on the MobileNetV2 neural network, adapted, for the emotion recognition task, using transfer learning techniques and the well-known facial expression dataset AffectNet.

The experiments have been run on a subset of the correctly classified image from the validation set. 10 images have been randomly selected, for each class for a total of 80 images. Moreover, three different experimental configurations have been defined, based on the number of filters applied to the input image: we used sequences of length equal to 3, 4, and 5. The filters' parameters *intensity* α and *strength* s are initialized with default values equal to 1.

Extending the concept of transfer learning, we decided to use hyperparameters that have been found for other problems [13,33]. As the results presented in Section 6.3 demonstrate, this proved to be a successful strategy as it allowed us to save time and computational effort without loss in protection effectiveness. Thus, we have chosen the following setup: for the outer algorithm, a population size = 10, mutation probability = 0.5, and 10 generations; the population size of the inner algorithm has been fixed to 5, and the number of generations to 3.

6.2. Evaluation

The system effectiveness is evaluated by the overall emotion attack success rate defined as:

$$EASR(X, X^*) = \frac{1}{n} \sum_{i=0}^{n} F(x_i) \neq F(x_i^*), \tag{4}$$

where n is the dataset size, and x_i and x_i^* are images from dataset X and the corresponding modified dataset X^*. We chose this measure as the standard evaluation measure for adversarial attacks to measure the percentage of images in the dataset for which the emotion recognizer fails the classification.

6.3. Results and Generated Images

The experimental results show that the algorithm can reach a high attack success rate EASR. More specifically, it achieves 91.25%, 93.75%, and 96.25% when using 3, 4, and 5 filters, respectively.

In Figures 1–3 the confusion matrices obtained by the three experiments are shown.

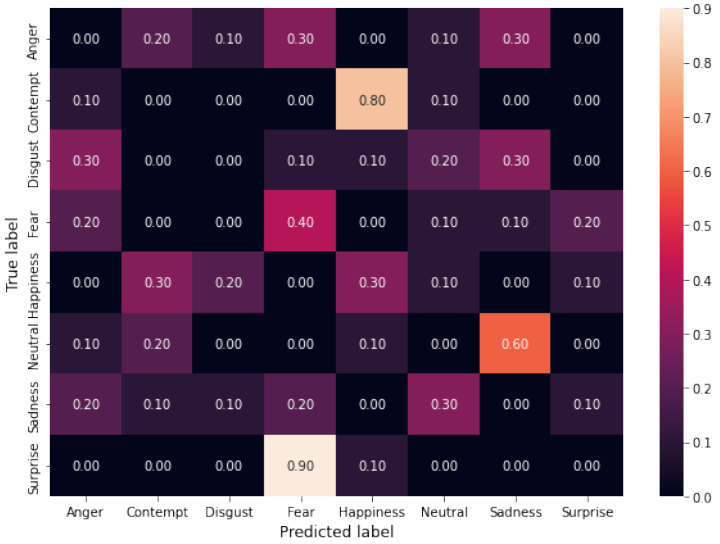


Figure 1. Confusion matrix from the results of the attack with three filters.

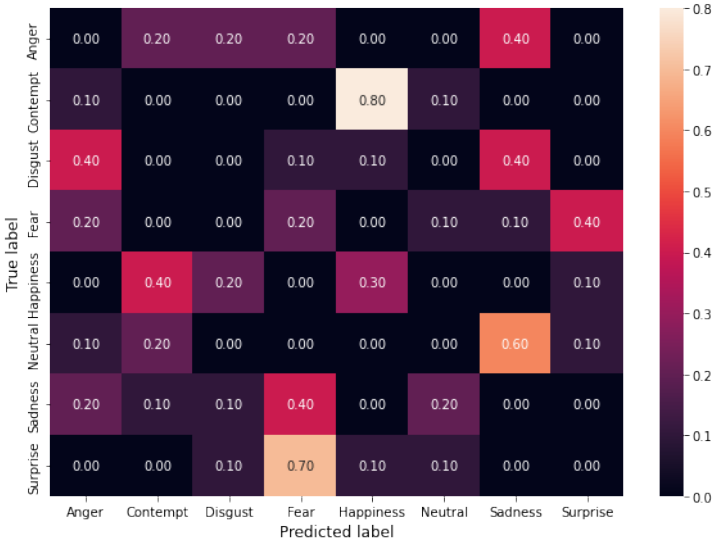


Figure 2. Confusion matrix from the results of the attack with four filters.

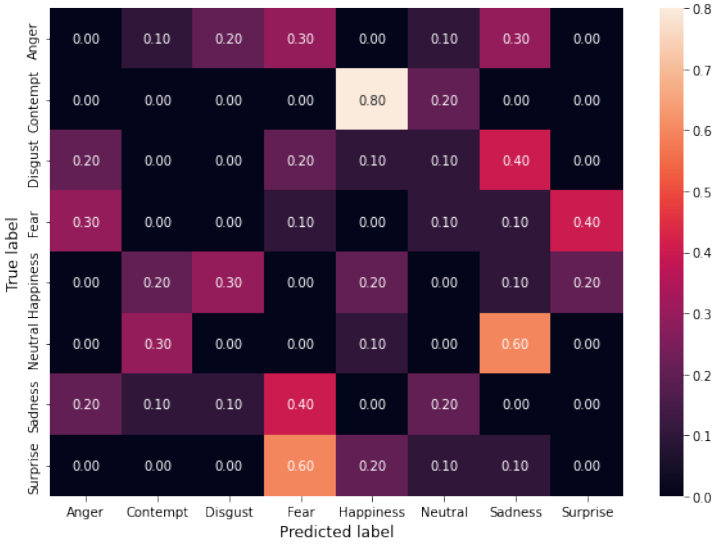
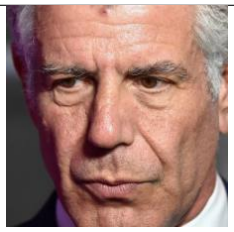
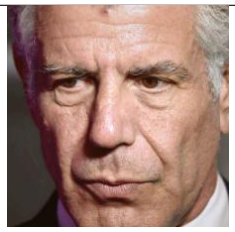
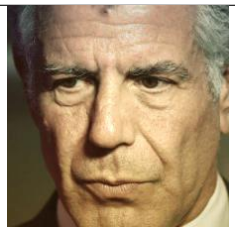
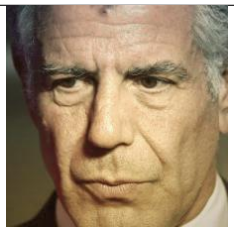


Figure 3. Confusion matrix from the results of the attack with five filters.

Confusion matrices allow evaluating the error distribution among classes. We can note that an increase of the length of filter sequences corresponds to increasing EASR and that the only classes that maintain some correct classifications are *fear* and *happiness* while all the others show an EASR of 100%. We ca also note that for the classes *Contempt*, *Neutral* and *Surprise* we obtained a shift (a number of errors greater than 50%) towards another class *Contempt* → *Happiness*, *Neutral* → *Sadness* and *Surprise* → *Fear*, while for the other five classes the errors are quite-uniformly distributed among the other classes.

Table 1. Examples of adversarial samples: the first column reports the original image and original classification. Columns 2–4 show the adversarial images with their classification. We can notice how the adversarial attack changes the automated emotion recognition without disrupting the image appearance.

Original	3 filters	4 filters	5 filters
			
surprise	fear	fear	fear
			
happiness	contempt	contempt	disgust
			
anger	sadness	sadness	sadness
			
happiness	contempt	contempt	contempt

We have also studied the impact of the filters on the images. In Table 1, some examples are shown. For each original image in the first column, the results obtained for sequences of 3, 4, and 5 filters are reported. We can note that the algorithm can produce natural-looking and artifacts free adversarial samples. This effect is due to the uniform application of the filters across the entire image, and the controlled perturbations through the SSIM index.

Moreover, to have a global view of the impact in terms of SSIM index, we have analyzed the values of the index for all the attacking images. In Figure 4, the distributions

of the SSIM values for the images produced by sequences of 3, 4 and 5 filters are shown. We can observe that, for most of the images, the scores are remarkably low, and only for very few cases, they reach values above 0.3.

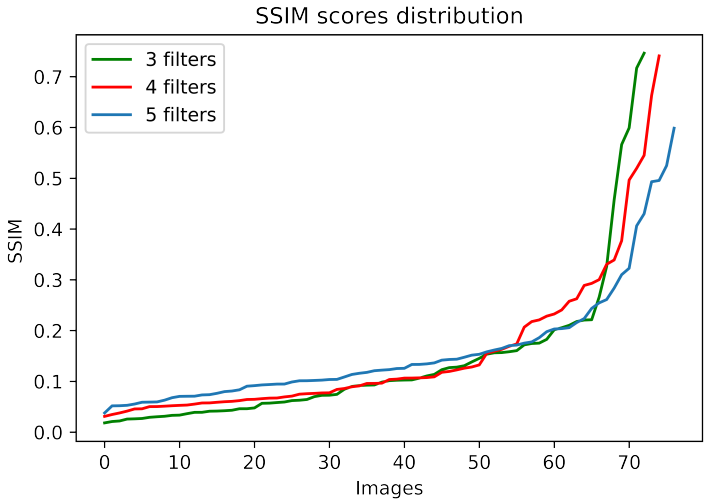


Figure 4. SSIM values distributions for the attacking images produced by sequences of 3, 4 and 5 filters.

We can also observe no significant differences among the three versions: users can choose according to their necessities, preferring a less/more modified image at the expense of the effectiveness of the protection.

7. Conclusions and Future Work

In a continuously evolving AI world, the most common applications of facial emotion recognition are prone to induce potentially unethical manipulation bias for the user's preferences. From the security point of view, deep learning models can be easily fooled by adversarial attacks, with data intentionally modified to avoid unauthorized information extraction by software. We show that combining a multi-objective evolutionary approach for AA with a per-instance approach allows the discovery of personalized sequences of filters having different image-specific characteristics, which can filter out the emotional features for the prying software. Composing and parametrizing popular image-enhancing Instagram filters on video frames or photos of human faces, the sequence is optimized by a nested-evolutionary algorithm, not requiring any information about the model's parameters or gradient values. Applying these protecting filters to any image we obtain a series of other images from which emotional information extraction is more difficult, while the transformation is transparent to the human perception, still keeping a natural look and their privacy-preserving emotional features. After a series of preliminary tests to have the best trade-off between computation efforts, time, and attacking effectiveness, achieving 91.25%, 93.75%, and 96.25% when using 3, 4, and 5 filters, respectively. The only classes that maintain some correct classifications are *fear* and *happiness*, all the others show an emotion adversarial attack success rate of 100%. Moreover, the algorithm can produce natural-looking and artifacts-free adversarial samples by applying the filters across the entire image and controlling perturbations through the structure similarity index.

Author Contributions: Conceptualization, V.F., A.M. and V.P.; methodology, V.P. and V.F.; software, G.B. and A.E.B.; validation, V.F., A.M., G.B., V.P. and A.E.B.; formal analysis, V.F., A.M. and V.P.; investigation, V.F., A.M., G.B., V.P. and A.E.B.; writing—original draft preparation, V.P., V.F., A.M., G.B. and A.E.B.; writing—review and editing, V.P. and V.F.; supervision, V.F., A.M. and V.P.; funding acquisition, V.F. All authors have read and agreed to the published version of the manuscript.

Funding: The APC was funded by International Workshops on Affective Computing and Emotion Recognition ACER/EMORE 2021.

Institutional Review Board Statement: The proposed research follows the terms of institutional commitments and regulations, applicable law, and standards of professional conduct and practice.

Informed Consent Statement: Not applicable.

Data Availability Statement: Source code is publicly available at https://github.com/Ellyuca/AGV-Project, accessed on 30 December 2021.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Picard, R.W. Affective Computing: Challenges. *Int. J. Hum.-Comput. Stud.* **2003**, *59*, 55–64. [CrossRef]
2. Gervasi, O.; Franzoni, V.; Riganelli, M.; Tasso, S. Automating facial emotion recognition. *Web Intell.* **2019**, *17*, 17–27. [CrossRef]
3. Sagonas, C.; Tzimiropoulos, G.; Zafeiriou, S.; Pantic, M. A Semi-automatic Methodology for Facial Landmark Annotation. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Portland, OR, USA, 23–28 June 2013; pp. 896–903.
4. Kazemi, V.; Sullivan, J. One millisecond face alignment with an ensemble of regression trees. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1867–1874.
5. Curumsing, M.K.; Fernando, N.; Abdelrazek, M.; Vasa, R.; Mouzakis, K.; Grundy, J. Emotion-oriented requirements engineering: A case study in developing a smart home system for the elderly. *J. Syst. Softw.* **2019**, *147*, 215–229. [CrossRef]
6. Franzoni, V.; Biondi, G.; Perri, D.; Gervasi, O. Enhancing Mouth-Based Emotion Recognition Using Transfer Learning. *Sensors* **2020**, *20*, 5222. [CrossRef] [PubMed]
7. Generosi, A.; Ceccacci, S.; Mengoni, M. A deep learning-based system to track and analyze customer behavior in retail store. In Proceedings of the 2018 IEEE 8th International Conference on Consumer Electronics-Berlin (ICCE-Berlin), Berlin, Germany, 2–5 September 2018; pp. 1–6.
8. Gorrini, A.; Crociani, L.; Vizzari, G.; Bandini, S. Stress estimation in pedestrian crowds: Experimental data and simulations results. *Web Intell.* **2019**, *17*, 85–99. [CrossRef]
9. Xing, Y.; Hu, Z.; Huang, Z.; Lv, C.; Cao, D.; Velenis, E. Multi-Scale Driver Behaviors Reasoning System for Intelligent Vehicles Based on a Joint Deep Learning Framework. In Proceedings of the 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Toronto, ON, Canada, 11–14 October 2020; pp. 4410–4415.
10. Ferrara, E.; Yang, Z. Quantifying the effect of sentiment on information diffusion in social media. *PeerJ Comput. Sci.* **2015**, *1*, e26. [CrossRef]
11. D'Errico, F.; Poggi, I. "Humble" Politicians and Their Multimodal Communication. In Proceedings of the Computational Science and Its Applications—ICCSA 2017, Trieste, Italy, 3–6 July 2017; Part III, Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2017; Volume 10406, pp. 705–717. [CrossRef]
12. Carpenter, J. The Quiet Professional: An Investigation of US Military Explosive Ordnance Disposal Personnel Interactions with Everyday Field Robots. Ph.D. Thesis, University of Washington, Washington, DC, USA, 2013.
13. Baia, A.E.; Di Bari, G.; Poggioni, V. Effective Universal Unrestricted Adversarial Attacks Using a MOE Approach. In Proceedings of the EvoAPPS 2021, Virtual Event, 7–9 April 2021.
14. Baia, A.E.B.; Milani, A.; Poggioni, V. Combining Attack Success Rate and Detection Rate for effective Universal Adversarial Attacks. In Proceedings of the ESANN 2021, online event, 6–8 October 2021.
15. Moosavi-Dezfooli, S.M.; Fawzi, A.; Frossard, P. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2574–2582.
16. Shamsabadi, A.S.; Oh, C.; Cavallaro, A. Edgefool: An Adversarial Image Enhancement Filter. In Proceedings of the ICASSP 2020, Barcelona, Spain, 4–8 May 2020.
17. Shahin Shamsabadi, A.; Sanchez-Matilla, R.; Cavallaro, A. ColorFool: Semantic Adversarial Colorization. In Proceedings of the CVPR 2020, Virtual, 14–19 June 2020.
18. Carlini, N.; Wagner, D. Towards Evaluating the Robustness of Neural Networks. In Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22–26 May 2017; pp. 39–57.
19. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. *arXiv* **2015**, arXiv:1412.6572.
20. Kurakin, A.; Goodfellow, I.J.; Bengio, S. Adversarial examples in the physical world. *arXiv* **2017**, arXiv:1607.02533.
21. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.J.; Fergus, R. Intriguing properties of neural networks. *arXiv* **2013**, arXiv:1312.6199.
22. Moosavi-Dezfooli, S.M.; Fawzi, A.; Fawzi, O.; Frossard, P. Universal Adversarial Perturbations. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 86–94.
23. Hayes, J.; Danezis, G. Learning universal adversarial perturbations with generative models. In Proceedings of the 2018 IEEE Security and Privacy Workshops (SPW), San Francisco, CA, USA, 24 May 2018; pp. 43–49.

24. Mopuri, K.R.; Ganeshan, A.; Babu, R.V. Generalizable data-free objective for crafting universal adversarial perturbations. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 2452–2465. [CrossRef]
25. Reddy Mopuri, K.; Krishna Uppala, P.; Venkatesh Babu, R. Ask, acquire, and attack: Data-free uap generation using class impressions. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 19–34.
26. Bae, H.; Jang, J.; Jung, D.; Jang, H.; Ha, H.; Lee, H.; Yoon, S. Security and privacy issues in deep learning. *arXiv* **2018**, arXiv:1807.11655.
27. Shokri, R.; Shmatikov, V. Privacy-preserving deep learning. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver, CO, USA, 12–16 October 2015; pp. 1310–1321.
28. Mireshghallah, F.; Taram, M.; Vepakomma, P.; Singh, A.; Raskar, R.; Esmaeilzadeh, H. Privacy in deep learning: A survey. *arXiv* **2020**, arXiv:2004.12254.
29. Liu, Y.; Zhang, W.; Yu, N. Protecting Privacy in Shared Photos via Adversarial Examples Based Stealth. *Secur. Commun. Netw.* **2017**, *2017*, 1897438. [CrossRef]
30. Liu, B.; Ding, M.; Zhu, T.; Xiang, Y.; Zhou, W. Using Adversarial Noises to Protect Privacy in Deep Learning Era. In Proceedings of the 2018 IEEE Global Communications Conference (GLOBECOM), Abu Dhabi, United Arab Emirates, 9–13 December 2018; pp. 1–6.
31. Xue, M.; Sun, S.; Wu, Z.; He, C.; Wang, J.; Liu, W. SocialGuard: An Adversarial Example Based Privacy-Preserving Technique for Social Images. *arXiv* **2020**, arXiv:2011.13560.
32. Sánchez-Matilla, R.; Li, C.; Shamsabadi, A.S.; Mazzon, R.; Cavallaro, A. Exploiting Vulnerabilities of Deep Neural Networks for Privacy Protection. *IEEE Trans. Multimed.* **2020**, *22*, 1862–1873. [CrossRef]
33. Arcelli, D.; Baia, A.E.B.; Milani, A.; Poggioni, V. Enhance while protecting: Privacy preserving image filtering. In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI-IAT '21), Melbourne, Australia, 14–17 December 2021.
34. Li, F.; Sun, Z.; Niu, B.; Guo, Y.; Liu, Z. SRIM Scheme: An Impression-Management Scheme for Privacy-Aware Photo-Sharing Users. *Engineering* **2018**, *4*, 85–93. [CrossRef]
35. Such, J.M.; Criado, N. Resolving Multi-Party Privacy Conflicts in Social Media. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 1851–1863. [CrossRef]
36. Xu, Y.; Price, T.; Frahm, J.M.; Monrose, F. Virtual U: Defeating Face Liveness Detection by Building Virtual Models from Your Public Photos. In Proceedings of the 25th USENIX Security Symposium (USENIX Security 16), Austin, TX, USA, 10–12 August 2016.
37. Wang, Z.; Bovik, A.; Sheikh, H.; Simoncelli, E. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef]
38. Papernot, N.; McDaniel, P.; Wu, X.; Jha, S.; Swami, A. Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks. In Proceedings of the 2016 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22–26 May 2016.
39. Akhtar, N.; Liu, J.; Mian, A. Defense Against Universal Adversarial Perturbations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3389–3398.
40. Xu, W.; Evans, D.; Qi, Y. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. *arXiv* **2018**, arXiv:1704.01155.
41. Mollahosseini, A.; Hasani, B.; Mahoor, M.H. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Trans. Affect. Comput.* **2019**, *10*, 18–31. [CrossRef]
42. Zhao, Z.; Liu, Z.; Larson, M. Adversarial Color Enhancement: Generating Unrestricted Adversarial Images by Optimizing a Color Filter. In Proceedings of the British Machine Vision Virtual Conference (BMVC), Virtual, 7–10 September 2020.
43. Wang, Y.; Wu, S.; Jiang, W.; Hao, S.; Tan, Y.a.; Zhang, Q. Demiguise Attack: Crafting Invisible Semantic Adversarial Perturbations with Perceptual Similarity. *arXiv* **2021**, arXiv:2107.01396.
44. Wang, L. A survey on IQA. *arXiv* **2021**, arXiv:2109.00347.
45. Xu, S.; Jiang, S.; Min, W. No-reference/Blind Image Quality Assessment: A Survey. *IETE Tech. Rev.* **2017**, *34*, 223–245. [CrossRef]
46. Zhai, G.; Min, X. Perceptual image quality assessment: A survey. *Sci. China Inf. Sci.* **2020**, *63*, 211301. [CrossRef]
47. Sun, Y.; Wu, C.; Zheng, K.; Niu, X. Adv-emotion: The facial expression adversarial attack. *Int. J. Pat. Recogn. Artif. Intell.* **2021**, *35*, 2152016. [CrossRef]
48. Sun, Y.; Yin, J.; Wu, C.; Zheng, K.; Niu, X. Generating facial expression adversarial examples based on saliency map. *Image Vis. Comput.* **2021**, *116*, 104318. [CrossRef]
49. Sharif, M.; Bhagavatula, S.; Bauer, L.; Reiter, M.K. Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, 24–28 October 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 1528–1540. [CrossRef]
50. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520. [CrossRef]
51. Ekman, P.; Friesen, W.V. A new pan-cultural facial expression of emotion. *Motiv. Emot.* **1986**, *10*, 159–168. [CrossRef]

52. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255. [CrossRef]
53. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]