Defending Emotional Privacy with Adversarial Machine Learning for Social Good

Shawqi Al-Maliki[*], Graduate Student Member, IEEE, Mohamed Abdallah*[*], Senior Member, IEEE, Junaid Qadir*[†], Senior Member, IEEE, Ala Al-Fuqaha*[*], Senior Member, IEEE*

[*] Information and Computing Technology (ICT) Division, College of Science and Engineering, Hamad Bin Khalifa University, Doha 34110, Qatar

[†] Department of Computer Science and Engineering, College of Engineering, Qatar University, Doha, Qatar

2023 International Wireless Communications and Mobile Computing (IWCMC) | 979-8-3503-3339-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/IWCMC58020.2023.10182780

***Abstract*—Protecting the privacy of personal information, including emotions, is essential, and organizations must comply with relevant regulations to ensure privacy. Unfortunately, some organizations do not respect these regulations, or they lack transparency, leaving human privacy at risk. These privacy violations often occur when unauthorized organizations misuse machine learning (ML) technology, such as facial expression recognition (FER) systems. Therefore, researchers and practitioners must take action and use ML technology for social good to protect human privacy. One emerging research area that can help address privacy violations is the use of adversarial ML for social good. Evasion attacks, which are used to fool ML systems, can be repurposed to prevent misused ML technology, such as ML-based FER, from recognizing true emotions. By leveraging adversarial ML for social good, we can prevent organizations from violating human privacy by misusing ML technology, particularly FER systems, and protect individuals' personal and emotional privacy. In this work, we propose an approach called *Chaining of Adversarial ML Attacks* (CAA) to create a robust attack that fools misused technology and prevents it from detecting true emotions. To validate our proposed approach, we conduct extensive experiments using various evaluation metrics and baselines. Our results show that CAA significantly contributes to emotional privacy preservation, with the fool rate of emotions increasing proportionally to the chaining length. In our experiments, the fool rate increases by 48% in each subsequent chaining stage of the chaining targeted attacks (CTA) while keeping the perturbations imperceptible ($\epsilon = 0.0001$).**

***Index Terms*—Evasion Attacks for Good, Emotional-Privacy Preservation, Robust Adversarial ML attacks.**

I. Introduction

Privacy is recognized as a fundamental human right by various international laws and regulations, including the Universal Declaration of Human Rights and the International Covenant on Civil and Political Rights. These laws and conventions establish that every individual has the right to privacy, including the right to control and protect their personal information and to be free of prying eyes and any unlawful or arbitrary interference. However, some organizations disregard established privacy regulations or lack transparency in their privacy practices, putting individuals' privacy at risk. In many cases, organizations violate privacy preservation by mishandling personal information, whether through failure to follow best practices for data protection or by intentionally selling private information to other entities.

Unscrupulous organizations often misuse machine learning (ML) technology, such as facial expression recognition (FER) systems, to violate privacy and cause social harm. Therefore, researchers and practitioners must take action and use ML technology for social good to protect individuals' privacy from such violations.

Adversarial ML for social good as an emerging research area can help address the concern of privacy violation in such situations. More specifically, evasion attacks can be repurposed to prevent misusing ML technology (e.g., ML-based FER) from recognizing true emotions. That motivates us to propose *Chaining of Adversarial ML Attacks* (CAA) to have a robust attack that fools a misused technology and prevents it from detecting the true emotions. We propose a generic chaining technique that can be applied to any adversarial ML attack, and in this work, we focus on chaining evasion attacks (also known as adversarial examples). We choose to investigate the Fast Gradient Sign Method (FGSM) attack as a representative example of an evasion attack because it is popular, simple, and efficient.

CAA is an auxiliary technique that can be added as a layer on top of any evasion attack to make it more robust. It provides the plain evasion attack with a higher success rate of fooling the target model. Beyond the conventional aspect of evasion attacks, our proposed work aims to contribute to the social good aspect where evasion attacks are enablers for human-centered applications. We define CAA as a sequence of adversarially generated attacks such that the output of the predecessor attack serves as an input to the successor attack with the aim of enhancing the robustness of the original attack. The chaining length is the number of involved stages in the chaining pipeline. We explore applying CAA on evasion attacks with the aim of getting a robust attack. The potential gains in robustness can be quantified by checking the impact of CAA on a few evaluation metrics (namely, the entropy of prediction, the popular L_p norm distances, and the fool rate) while considering the unchained attacks (i.e., plain evasion attacks) as the baseline.

In this work, emotions refer to facial expressions, and emotion recognition system (ERS) refers to (FER). Social good and human-centric are used interchangeably. Our proposed work is motivated by the ubiquitous situations (such as students in schools and universities as well as employees in private and public sectors) where identity recognition is not a concern but emotion is. That is, their identities should be recognized, while their emotions should remain private.

While adversarial ML is a rapidly growing research area [1], we are the first, according to the best of our knowledge, to propose a general augmenting technique that can improve the robustness of any evasion attack. While several works exist that utilize evasion attacks as Human Centric (HC) enablers, such as [2]–[10], there is a lack of research on utilizing evasion attacks for emotional-privacy preservation.

The salient contributions of this work are as follows.

- We propose augmenting plain evasion attacks with a chaining technique called CAA, which is a generic technique that can make any evasion attack more robust.
- We demonstrate how the proposed chaining technique CAA can be harnessed to use evasion attacks for applications aimed at social good. As a case study, we leverage CAA as a privacy-preserving enabler for emotion recognition systems.
- We conduct extensive experiments to validate the proposed chaining technique.

We organize the remaining sections of this paper as follows. Section II summarizes the related works. Section III illustrates and explains our proposed approach. The experiments and results are shown in Section IV. Section V concludes our work and presents future works.

II. Related Work

A. Evasion Attacks

Evasion attacks, also known as adversarial examples, are a type of adversarial ML attack in which imperceptible perturbations are added to original inputs to generate outputs that can fool ML models during inference [11], [12]. The most commonly used evasion attacks include L-BFGS [12], Fast Gradient Sign Method (FGSM) [13], Carlini and Wagner [14], DeepFool [15], projected gradient descent (PGD) [16], and auto-attacks [17]. These attacks are typically used to explore the limits of Deep Neural Networks (DNNs) and to highlight or leverage security concerns related to adversarial robustness.

In this paper, we propose an auxiliary technique (CAA) that can be applied to any evasion attack to make it more robust. In other words, an evasion attack, when complemented with CAA, has a higher success rate of fooling the target model. Beyond the conventional aspects of evasion attacks, our work aims to contribute to the social good aspect of adversarial attacks, specifically in the context of emotional-privacy preservation. We utilize FGSM as a representative evasion attack because it is simple, fast, efficient, and suitable for our use case. However, our proposed chaining technique is generic and can be applied to any adversarial ML attack. To the best of our knowledge, our work is the first to propose a generic augmenting technique that can be applied to any evasion attack to enhance its robustness.

B. Evasion Attacks for Good

Evasion attacks are conventionally used in contexts where the security of ML is a concern. Recently, utilizing evasion attacks as an enabler for human-centric applications has emerged as a new line of work in the area of adversarial ML. Examples of applications that utilize evasion attacks as a human-centric enabler are model reprogramming for data-efficient transfer learning [2], [3]; contrastive explanations and counterfactual examples [4]; model watermarking and fingerprinting [18], [19]; data cloaking for enhanced privacy and data security [5], [6], data augmentation for improving model generalization [7], [8], and robust text CAPTCHAs [9], [10]. The social-good-centered aspect of evasion attacks can be incorporated into further human-centered applications. In recent times, emotion recognition systems have become ubiquitous with less obvious regulations that protect the emotional privacy of people, but work focused on using evasion attacks for emotional-privacy preservation is lacking. Our work aims to plug this gap and we focus our proposed adversarial attack augmentation technique for enhancing emotional-privacy preservation.

III. System Model

This section presents the proposed scheme and the application scenarios. Fig. 1 shows our proposed scheme for handling emotion-privacy preservation. It comprises CAA and ERS.

A. Chaining of Adversarial ML Attacks (CAA)

CAA is a need-driven service that is activated in application scenarios where the existence of unauthorized ERS is a concern. In other words, CAA can be triggered when a user wants to preserve his emotions and deactivated otherwise. CAA is a pipeline of stages where the output of a previous stage is utilized as an input to the next stage. A chaining stage is a unit of generating the adversarial attack. It receives an image, utilizes a trained model for generating an adversarial attack corresponding to the received image, and then passes the attacked image to the next chain to repeat the same process (as illustrated in Algorithm 1). Our attack is with a white-box threat model where the trained model utilized for the chaining process is assumed to have the same architecture and parameters as the main prediction model of ERS (i.e., ERM). CAA can be applied as a targeted or untargeted evasion attack. The former requires specifying the label of a target, while the latter does not. In addition, a targeted attack is optimized to maximize the probability of the target label, while an untargeted attack is optimized to minimize the probability of the true label [20]. Fig. 1 illustrates the targeted evasion attacks as an example. Untargeted evasion attacks can be represented with the same illustration but without specifying in advance the target label. We keep it to the model to optimize the selection of the feasible label.

B. Emotion Recognition System (ERS)

Our proposed scheme is a need-driven service that is activated or deactivated based on the application scenario. ERS can be an authorized entity, so CAA is deactivated. It also can be an unauthorized component hidden within platforms authorized for other specific tasks rather than emotion recognition. Examples of such platforms are identity recognition or video conference platforms. In such suspected scenarios, our proposed scheme is activated. The main component of ERS

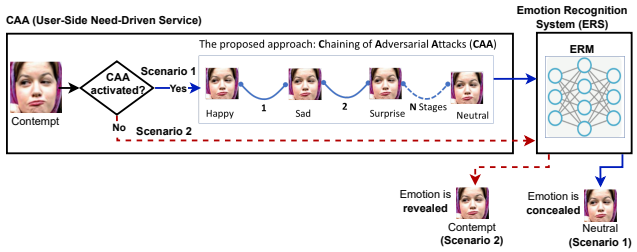


Fig. 1: Abstract view of the proposed CAA approach.

Algorithm 1 The Proposed Chaining Algorithm (CAA)

Input: Unattacked emotion (revealed emotion).
Output: Attacked for social good emotion (concealed emotion).

```
if CAA is activated then
    for chaining stage = 1, ..., N do
        Receive an emotion
        Generate an adversarial attack corresponding to the received emotion
        Pass the attacked emotion to the next chain
    end for
else
    Pass the emotion directly to ERS
end if
```

is the Emotion Recognition Model (ERM), which predicts the labels of the received emotions. The output of the last chaining stage in CAA is a robust evasion attack that is passed to ERM to fool it with a high probability of success. In other words, by utilizing CAA, we ensure that emotional privacy is preserved by concealing the true emotion. For example, as illustrated in Fig. 1, the true facial expression emotion was *Contempt*, and for preserving privacy right, this emotion is passed through stages of attacks to enhance the potential of concealing the true emotion. For example, the image is adversarially attacked to have faked label (*Happy*) and again attacked to have *Sad, Surprise*, and *Neutral* labels in the subsequent stages, respectively. For that, it fools the ERM easily and confidently. That is, it manages to preserve privacy and hide true facial expressions. We can conceal the true emotion and replace it with a specific class of emotions by making the target class of the last chaining stage the specific emotion we target. We explain the application scenarios of CAA in the next section.

C. Application Scenarios

Our proposed work is motivated to be used in scenarios where identity recognition is not a concern, but emotion recognition is. These scenarios can be the online video platforms used for online teaching or meeting in universities, schools, companies, or their corresponding metaverse representations. However, these platforms are unauthorized for emotion recognition. In such settings, ERS starts running and collecting emotions without the consent of concerned students and employees.

IV. Experiments and Results

A. Baselines and Performance Metrics

*1) **Baselines**:* The utilized baselines are the unchained (i.e., plain) targeted and untargeted FGSM attacks [13]. We refer to them as Plain Targeted Attack (PTA) and Plain Untargeted Attack (PUA). We configured the target label of PTA to be the label with the least predicted probability, as this is the most challenging target label. For Plain Untargeted Attack (PUA), by definition, we leave the target label unspecified.

*2) **Performance Metrics**:* We investigate the performance of our proposed approach on targeted and untargeted FGSM evasion attacks (details on evasion attack types are in Section III-A). When we apply our proposed chaining approach (CAA) to targeted attacks, we call them Chained Targeted Attacks (CTA). Likewise, when we apply CAA on untargeted attacks, we call them Chained Untargeted Attacks (CUA).

In this work, we are interested in evaluating the impact of CTA and CUA, compared to their corresponding baselines, on the entropy of model predictions, L_p norm distance metrics (L_0, L_2, L_∞), and the fool rate. Entropy measures the randomness in model predictions (denoted as $p(x)$) as shown in Equation 1).

$$H(x) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i). \tag{1}$$

For each investigated attack, we calculated the entropy of predictions as the average of the whole test dataset. L_0 counts the number of perturbed pixels in the attacked image, L_2 measures the distances between the clean and its corresponding attacked images, and L_∞ measures the maximum perturbation in all pixels of the image. Fool rate, on the other hand, is the ratio of the successfully performed attacks out of the overall attacks.

To ensure that the generated attacks are imperceptible to humans (to align with the definition of the evasion attacks [12]), for all investigated metrics, we performed the attacks using a representative small perturbation threshold 0.001. We also re-run the experiments on further smaller perturbations (e.g., 0.0001) to investigate the impact of our proposed approach on such smaller changes.

B. Experimental Setup

We investigate the feasibility of our proposed approach (CAA) by considering its impact on the entropy of model predictions, popular adversarial attacks related L_p norm distance metrics (L_0, L_2, L_∞), and the fool rate (details of these metrics on Section IV-A). We perform experiments on AffectNet [21]. It is a popular Facial Expression Recognition (FER) dataset comprising eight classes (Anger, Contempt, Disgust, Fear, Happy, Neutral, Sad, Surprise). We utilized the Adversarial-Robustness-Toolbox (ART) [22] to perform the attacks. ART is a library for various adversarial machine learning attacks, including evasion attacks, which is the focus of this work. As a representation attack strategy, we applied our proposed chaining approach to the popular FGSM attack [13]. Chaining can be applied to any adversarial attack, though.

We perform extensive experiments to answer the following research questions:

1) What is the impact of CAA on the entropy of model predictions?
2) What effect does CAA have on the distance metrics (L_0, L_2, L_∞)?
3) What is the impact of CAA on the fool rate?
4) What effect does CAA have on the investigated metrics when the perturbations go further smaller?

We answer these questions in the following section.

C. Experimental Results

*1) **The impact of chaining on the entropy of model predictions**:* As illustrated in Fig. 2, we observe that compared to the entropy of the two baselines (PTA and PUA), the associated entropy of the proposed chained attacks (CTA and CUA) keep increasing in each subsequent chain. The more chaining stages, the more entropy.

The ongoing increment of the entropy on the proposed chained attacks (CTA and CUA) can be interpreted as an ongoing increment in the uncertainty, which in turn means more potential for concealing the true emotion (i.e., a more robust attack).

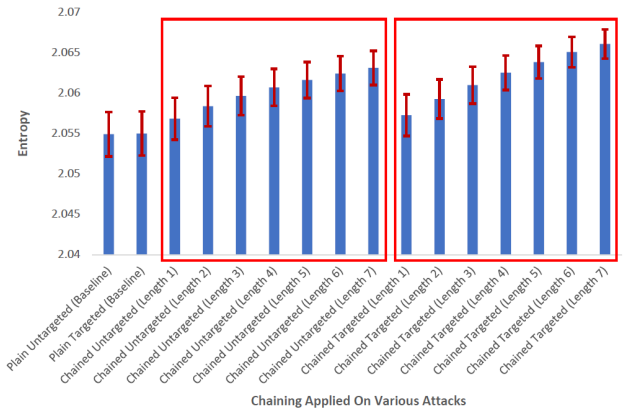
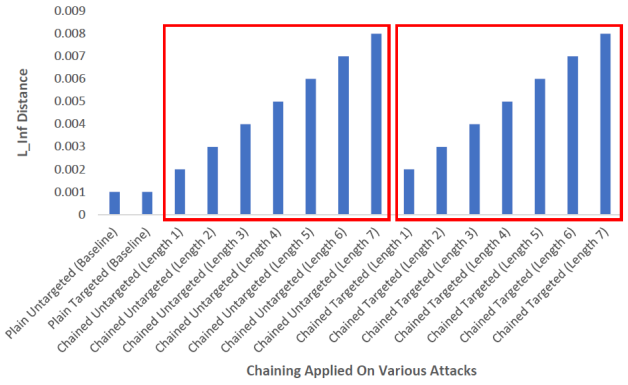


Fig. 2: Entropy Average with Various Attacks (ϵ = 0.001).

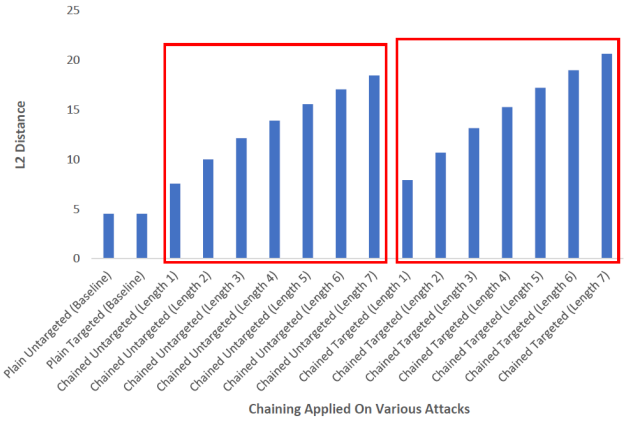
*2) **The effect of CAA on distance metrics** (L_∞, L_2, L_0):* By analyzing Fig. 3, we observe that for L_∞ distance (Fig. 3a), the distance increases with the added chain. The more chaining, the higher L_∞ distance. Each chain adds a new perturbation. So, the accumulative perturbation correlates with the perturbation value (ϵ) and the chaining length.

Regarding L_2 distance, Fig. 3b illustrates that the proposed chaining attacks (CUA and CTA) make the sequence of attacked inputs further away from the corresponding baselines (PUA and PTA). The higher distance between the original and attacked images, the more robust the attacked image is. That is, it can be used for emotional-privacy preservation with a high probability of success.

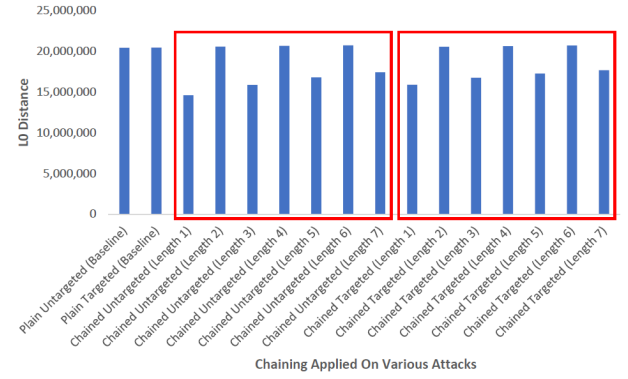
Concerning L_0 distance as shown in Fig. 3a, we notice that the Chain Targeted and Chain Untargeted (CUA and CTA) have the same pattern. One chain increase and subsequent decrease. One interpretation for the pattern of why one chain is increased and the subsequent chain is decreased is because one chain may cancel another. The Representative attack, FGSM, works by adding or subtracting a perturbation to/from the current values. That means that a newer chain adds perturbations by either adding or subtracting from the value of each pixel of the image of the previous chain. Thus, a newer chain either restores the original or changes the value of the pixel.



(a) L_∞ distance (ϵ = 0.001).



(b) L_2 distance (ϵ = 0.001).



(c) L_0 distance (ϵ = 0.001).

Fig. 3: The effect of CAA on the popular distance metrics used in adversarial ML attacks' literature.

*3) **The impact of CAA on the fool rate**:* Fig. 4 shows that in comparison to the established baselines (PUA and PTA), the fool rate of the corresponding proposed chained attacks keeps increasing with the subsequent chaining stages.

An increased fool rate indicates a stronger attack, which means a higher potential for emotional-privacy preservation. We observe that though both proposed chained attacks (CTA and CUA) highly correlated with chaining length, the chained untargeted attacks (CUA) outperform chained targeted attacks (CTA). This observation makes sense as the targeted attacks (PTA & CTA) are more challenging by definition. Targeted attacks are achieved by optimizing the probability of *all* predicted labels (except the true label) to be less than the probability of the true label. While the optimization in the untargeted attacks is done by minimizing *only* the predicted probability of the true label [13].

When the perturbation goes smaller, the ratio of the increased fool rate goes higher. That clearly indicates that our proposed chaining approach (CAA) shines when the perturbation goes smaller. That makes CAA appealing because it's shining in smaller perturbations means that the attacked images are more imperceptible to humans.

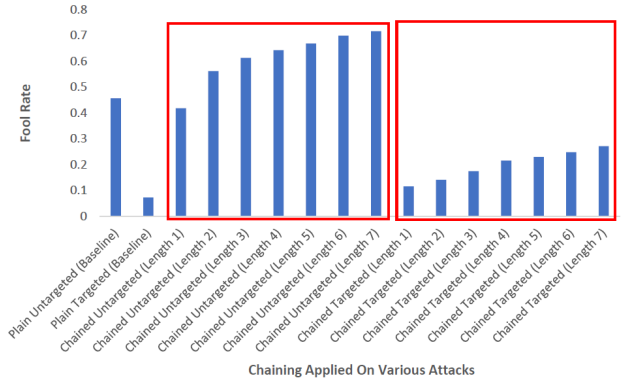


Fig. 4: The impact of CAA on the model's fool rate for various attacks.

*4) **The effect of CAA on all evaluation metrics when the perturbation value is small**:* When the perturbation value is 0.001, the average of the incremental ratio of the fool rate is 0.1090 in the case of CUA and 0.3604 in the case of CTA. However, as illustrated in Fig. 5, whenever the perturbation becomes further imperceptible, such as $\epsilon = 0.0001$, the average incremental ratio of the fool rate becomes higher. Precisely, 0.3925 in the case of CUA and 0.4833 in the case of CTA. That indicates that our proposed chaining shines with smaller perturbation values. That is an advantage because smaller perturbations mean high imperceptibility of the attack.

With regards to the entropy, the average of the incremental ratio is 0.0005 and 0.0007 when the perturbation is 0.001 for CUA and CTA, respectively. However, when the perturbation goes smaller (e.g., 0.0001), the average incremental ratio of the entropy becomes unchanged (0.0001) for both, CUA and CTA.

The ratio of L_∞ distance does not change for smaller perturbations. It is increasing in each subsequent chain (CTA or CUA) with a ratio of 100% regardless of the perturbation value. That is an intuitive observation because L_∞ measures the degree of perturbations. The more perturbation, the more L_∞ distance.

Concerning the ratio of L_2 distance, the average of the incremental ratio is (0.4001, 0.4671) for 0.001 perturbation of the CUA and CTA, respectively. Whenever the perturbation goes smaller (e.g., 0.0001), the average incremental ratio of the L_2 distances becomes higher (0.8682, 0.8975) for CUA and CTA. It is obvious that the L_2 distance almost gets doubled when the perturbation is smaller, which is another evidence that implies that our proposed approach shines with smaller perturbations.

Regarding the effect of going with a smaller perturbation on the L_0 distance, we observe that the smaller perturbation, the lesser the L_0 distances. For example, (0.0230, 0.0145) are the associated L_0 distances for CUA and CTA when the perturbation is 0.001. Whenever the perturbation goes smaller (e.g., 0.0001), the L_0 distance becomes 0.0013 and 0.0008, which is smaller. That is interpreted by the number of changed pixels. L_0 distance captures less number of changed pixels whenever the perturbation goes smaller.

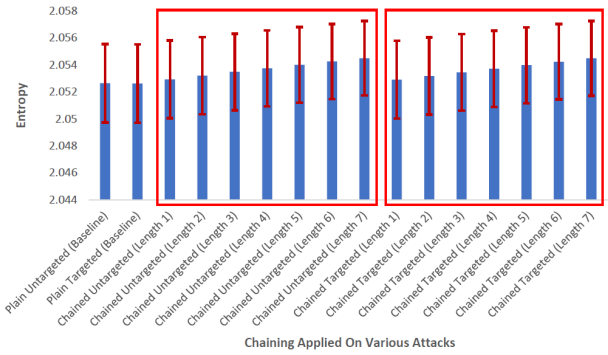
D. Discussion and Insights

Our experiments confirm that the proposed chaining approach (CAA) has a high impact on the entropy of prediction, fool rate, L_2 distance, and L_∞. They are all highly correlated. The more entropy, the more fool rate, and the more L_2 and L_∞ distances. The observed correlation is invaluable in emphasizing the impact of CAA to augment the robustness of the generated attacks and consequently the emotional privacy preservation.

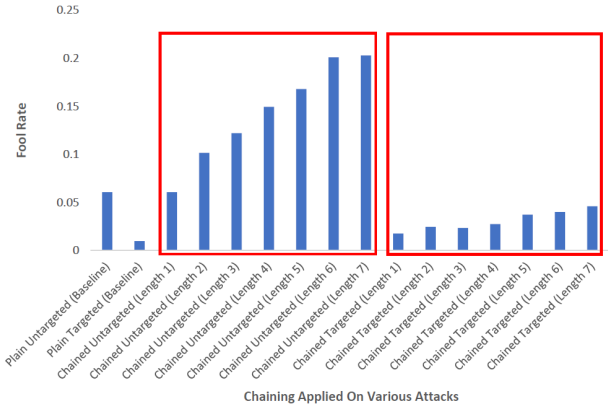
For real-world applications and during inference time, measuring the fool rate is challenging. It requires having ground truth labels, which are financially and operationally expensive. Similarly, measuring L_2 and L_∞ distances at inference time is challenging as that necessitates the existence of the original image (before the adversary perturbed it). On the other hand, measuring the entropy at inference time is feasible because it depends on the predictions and not on the ground truth labels. Besides, it does not require the original images or the ground truth labels. Since we notice that entropy is highly correlated with the remaining investigated metrics, it can be used as a feasible and reliable proxy for quantifying the other infeasibly quantified metrics at inference time. Our proposed chaining technique (CAA) shines when perturbation goes smaller, which makes CAA appealing. Smaller perturbations mean more imperceptibility of the attack. For that, with our proposed chaining (CAA), emotions can be concealed with a high probability of success.

V. Conclusions and Future Work

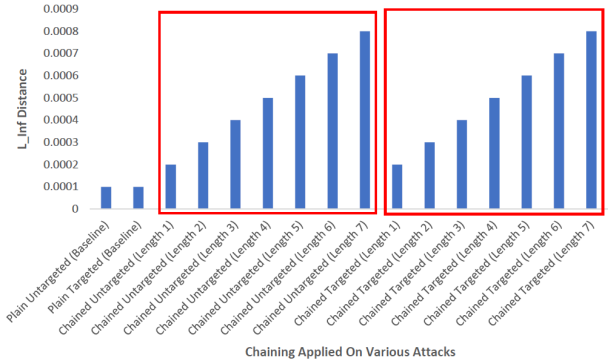
In this work, we propose a Chaining of Adversarial ML (CAA). A generic technique that can complement any evasion attack to enhance its robustness to contribute effectively as an enabler for evasion attacks for social good applications. As a showcase, we apply our proposed approach as a privacy-preserving enabler for emotion recognition systems such that privacy is protected by the concerned users and not by their organizations, which may incidentally or intentionally violate users' privacy. In other words, privacy assurance and



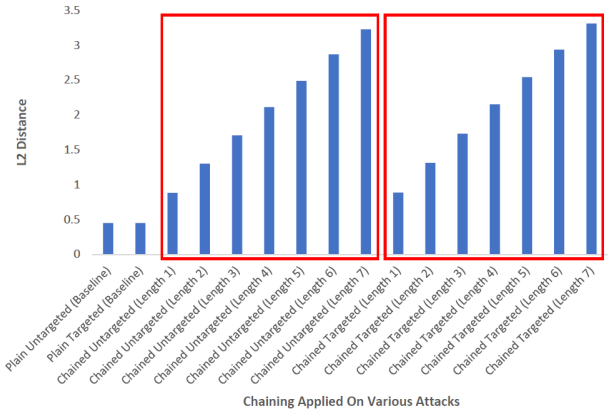
(a) Entropy Average when $\epsilon = 0.0001$.



(b) Fool rate ($\epsilon = 0.0001$).



(c) L_∞ distance ($\epsilon = 0.0001$).



(d) L_2 distance ($\epsilon = 0.0001$).

Fig. 5: The impact of CAA when the perturbation goes smaller.

enforcement are switched to the concerned users to have the needed upper hand. In our future work, we plan to investigate the impact of CAA on attack reversibility.

REFERENCES

[1] N. Carlini. A Complete List of All (arXiv) Adversarial Example Papers . [Online]. Available: https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html

[2] G. F. Elsayed, I. Goodfellow, and J. Sohl-Dickstein, "Adversarial reprogramming of neural networks," *arXiv preprint arXiv:1806.11146*, 2018.

[3] Y.-Y. Tsai, P.-Y. Chen, and T.-Y. Ho, "Transfer learning without knowing: Reprogramming black-box machine learning models with scarce data and limited resources," in *International Conference on Machine Learning*. PMLR, 2020, pp. 9614–9624.

[4] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das, "Explanations based on the missing: Towards contrastive explanations with pertinent negatives," *Advances in neural information processing systems*, vol. 31, 2018.

[5] S. Shan, E. Wenger, J. Zhang, H. Li, H. Zheng, and B. Y. Zhao, "Fawkes: Protecting privacy against unauthorized deep learning models," in *29th USENIX security symposium (USENIX Security 20)*, 2020, pp. 1589–1604.

[6] V. Cherepanova, M. Goldblum, H. Foley, S. Duan, J. Dickerson, G. Taylor, and T. Goldstein, "Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition," *arXiv preprint arXiv:2101.07922*, 2021.

[7] C.-Y. Hsu, P.-Y. Chen, S. Lu, S. Liu, and C.-M. Yu, "Adversarial examples can be effective data augmentation for unsupervised machine learning," in *AAAI Conference on Artificial Intelligence*, 2022.

[8] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[9] R. Shao, Z. Shi, J. Yi, P.-Y. Chen, and C.-J. Hsieh, "Robust text captchas using adversarial examples," *arXiv preprint arXiv:2101.02483*, 2021.

[10] C. Shi, X. Xu, S. Ji, K. Bu, J. Chen, R. Beyah, and T. Wang, "Adversarial captchas," *IEEE Transactions on Cybernetics*, 2021.

[11] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*. Springer, 2013, pp. 387–402.

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

[13] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[14] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 39–57.

[15] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2574–2582.

[16] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *arXiv preprint arXiv:1607.02533*, 2016.

[17] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *ICML*, 2020.

[18] S. Wang, X. Wang, P.-Y. Chen, P. Zhao, and X. Lin, "Characteristic examples: High-robustness, low-transferability fingerprinting of neural networks." in *IJCAI*, 2021, pp. 575–582.

[19] A. Sablayrolles, M. Douze, C. Schmid, and H. Jégou, "Radioactive data: tracing through training," in *International Conference on Machine Learning*. PMLR, 2020, pp. 8326–8335.

[20] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, 2018, pp. 2154–2156.

[21] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[22] M.-I. Nicolae, M. Sinn, M. N. Tran, B. Buesser, A. Rawat, M. Wistuba, V. Zantedeschi, N. Baracaldo, B. Chen, H. Ludwig, I. Molloy, and B. Edwards, "Adversarial robustness toolbox v1.2.0," *CoRR*, vol. 1807.01069, 2018. [Online]. Available: https://arxiv.org/pdf/1807.01069