Adversarial Attacks and Defences for Infrared Deep Learning Systems

Flaminia Spasiano
Sapienza University
Rome, Italy

Gabriele Gennaro
Leonardo Labs
Rome, Italy

Simone Scardapane
Sapienza University
Rome, Italy

***Abstract*—This paper studies adversarial attacks and defences against deep learning models trained on *infrared* data to classify the presence of humans and detect their bounding boxes. Firstly, we study the effectiveness of the Projected Gradient Descent (PGD) adversarial attack against Convolutional Neural Networks (CNNs) trained on infrared data, and the effectiveness of adversarial training as a possible defense against the attacks. Secondly, we study the response of an object detection model trained on infrared images under adversarial attacks. In particular, we propose and empirically evaluate two attacks: one classical attack from the literature on object detection, and a new hybrid attack which exploits a common CNN base architecture of the classifier and a detector. We show for the first time that adversarial attacks weaken the performance of classification and detection models trained on infrared images only. We also prove that the defense adversarial training optimized for the infinity norm increases the robustness of different classification models trained on infrared data.**

***Index Terms*—Deep learning, infrared, adversarial attack, adversarial training**

I. Introduction

Deep neural networks-based (DNNs) systems have found wide application in many areas and, therefore, have become enticing targets for adversaries, such as hackers or criminal/government organizations, which may seek to 'break' the systems by exploiting their brittleness to out-of-sample inputs and malicious inputs [1]. Thus, it is critical to learn and employ efficient methods to defend trained models, especially when the systems are used for monitoring or surveillance of areas of interest. Nowadays, a vast literature exist on possible attacks and defences against trained computer vision systems [2], but the vast majority has focused on systems developed for RGB images in the visible spectrum, due to the wide availability of data and pre-trained models. In this paper, we focus instead on adversarial systems for infrared cameras, which are commonly used in practice but have received less attention from the adversarial machine learning literature.

The infrared is part of the electromagnetic spectrum and it is imperceptible to the human eye. An infrared 'image' is different from a RGB image: it has a single channel instead of three (red, green and blue), it has a grey-scale coloring, and it contains nor texture or the luminescence of the subject. These images can be highly informative and reveal hidden (in the visible spectrum) properties of a subject. Therefore, in many applications we may look beyond the visible-light range of wavelengths, e.g., thermal imaging [3] [4], monitoring the earth through satellites [5], and recognition and detection of objects [6] [7]. In particular, these last two applications are the most subject to adversarial attacks, for example to dodge surveillance cameras, and pass authentication systems.

In the case of the visible spectrum, adversarial examples are proven to exist against neural networks, and they are also shown to be transferable from one network to another [1]. On the other hand, regarding the infrared spectrum, only last year Edwards and Rawat [8] studied the effectiveness of adversarial attacks against DNNs trained on simulated infrared images and the effectiveness of adversarial training on such models. Prior to their work, the effectiveness of adversarial attacks on deep learning systems trained on 1 channel from the non-visible spectrum was shown by no published research. Recent studies demonstrated how practical attacks can be really effective against identity verification systems trained on an infrared dataset [9], but non-physical attacks on classification/detection models trained on infrared datasets are still a topic to be explored.

In this paper we show that adversarial attacks against both classification models and detection models trained on infrared data works and lower the performance of such models. In particular, the Projected Gradient Descent (PGD) attack is capable of decreasing the accuracy of simple and complex classification networks trained on infrared images, and we show that its strength can be significantly mitigated through an adversarial training defence. We also prove that an object detection model trained on infrared images is vulnerable to adversarial attacks of different natures. Our results demonstrate that deep learning systems trained on infrared data are sensitive to white box attacks and therefore not secure unless specific mitigation measures are adopted (e.g., adversarial defence).

II. Dataset

In this paper we consider only infrared images in the thermal wavelength interval. The dataset was constructed by the ASL dataset [10] and it is composed of different sequences of infrared images shot with a resolution of 324×256 pixels by a FLIR Tau 320 camera, a thermal imaging camera sensible to the Longwave Infrared range (LWIR). The chosen dataset contains 4381 thermal infrared images involving mainly humans. The authors provided annotated ground truth for all object in each images. The LWIR wavelength is the most commonly used form of infrared technology because LWIR

Fig. 1. Examples images from the dataset [10].

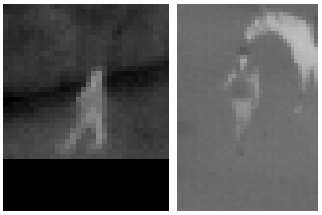

Fig. 2. Examples of cropped processed images for the classification task.

Fig. 3. Examples images for the human detection task.

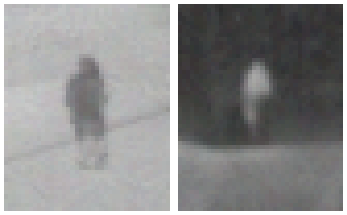

Fig. 4. Some examples of adversarial images created with projected gradient descent attack

cameras detect radiated temperature (for example [11]). In Fig. 1 we show two images randomly sampled from the dataset.

A. Classification Data

In order to have appropriate data for the task of image classification, we crop each image in the following way: instead of dealing with the whole picture, each object is cropped with some margin to include background and other shapes, as depicted in Fig. 2. Random crops of the background are included to balance the dataset. The dataset was then pre-processed with pixel normalization and fixing the common image size by eventual padding. The final dataset contains 13860 images being composed by actual and clear figures of humans through infrared vision.

B. Detection Data

The input data used in the task of object detection is instead the original dataset. In total there are 4072 images divided among test and train dataset. Two randomly sampled images are shown in Fig. 3.

III. Attacks for infrared classification systems

To assess the hardness of the problem and to have a comparison, the same structure of the experiment conducted by [8] was adopted. Two architectures with different complexity are employed, in order to be sure the complexity does not cause change in accuracy after adversarial training. One pretrained ResNet model with benchmarks on other RGB datasets, and one simple 8-layered CNN (AlexNet) (trained from the ground up with random weights at initialization) are considered. Both architectures are tested on infrared imagery and then tested on adversarial examples generated with the Projected Gradient Descent attack. After adversarial training again both models are tested on clean and perturbed data from the test set and the adversarial test set. The adversarial set is built in the following way: a random subset is taken out of the test set, then each image is attacked with the Projected Gradient Descent attack and added to the adversarial set. generated with this process are instead shown in Fig. 4.

IV. Attacks for infrared detection systems

In the second part of the experimental results, we study instead how an object detection model reacts under two particular attacks, and we evaluate its performance under these scenarios. The first attacks is an hybrid attack which exploits a common CNN base architecture of the classifier and a detector. The second attack is a classical attack from the object detection attacks literature.

Detectron2 [12] is the chosen object detection library, as it provides state-of-the-art detection and segmentation performance on a number of standard benchmarks. Also currently there are no published researches of attacks against model implemented through Detectron2 and trained on infrared data. The model we use is a Faster R-CNN [13] with a ResNet-50 backbone and Feature Pyramid Networks (FPN) region proposal method [14]. The metric used for all comparisons and evaluations is the the average precision metric. For our experiments, Faster R-CNN is pretrained on the COCO dataset, an RGB image dataset. After initializing the configuration parameters, Faster R-CNN is trained on the chosen dataset of infrared images. In Table I we collect the baseline performance of the model on the COCO dataset, and the average precision on the target dataset after training on infrared images.

A. First proposed object detection attack

Since adversarial attacks for object detection models need to mislead not only the classification results but the whole structure (i.e., bounding boxes), attacking an object detection model is much harder than attacking an image classifier.

In our first proposed algorithm for attacking a detection model, the aim is to prove that by having the knowledge that a classification model and a detector share the same base CNN, it is possible to craft adversarial examples that cause

TABLE I
BENCHMARK PERFORMANCE FOR FASTER R-CNN, ON THE ORIGINAL RGB IMAGES AND ON THE CHOSEN INFRARED DATASET.

Dataset	Box Average Precision
COCO RGB	40.2 %
Infrared	63.3 %

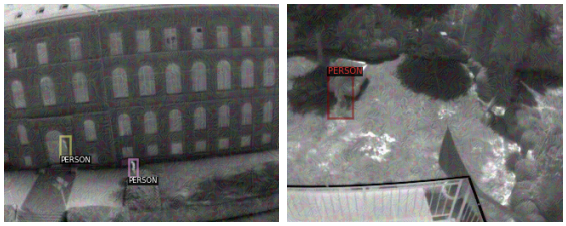

Fig. 5. Some examples of adversarial images created with projected gradient descent attack against the detection model

inefficiency on the classifier as well as on the detector. This implies the possibility for the existence of an attacker able to transfer an attack from a simple model to a much more complex model without any effort than the creation of this adversarial example (which is negligible).

The models employed are the previous classification ResNet-50 architecture trained on infrared data and Faster R-CNN with the ResNet backbone. These two different approaches share the same CNN base. The idea of attack transferability takes advantage of the fact that both models share the same base architecture, they are pretrained on RGB images, and they are retrained on infrared images belonging to the same dataset. This implies that an adversarial example generated to fool the ResNet classification model, might also 'fool' the ResNet backbone of Faster R-CNN, causing a deviation to the wrong output in the process of feature extraction.

More specifically, we consider the following pipeline for attacking the object detection model:

1. Firstly, the previous infrared-trained ResNet classifier generates an adversarial dataset using data coming from the Object-Detection dataset. In other words the previous ResNet classifier exploits the Projected Gradient Descent on object detection data to create fooling images against ResNet. A couple of examples are depicted in Fig. 5.
2. Secondly, the adversarial dataset is used in input to Faster R-CNN.

Fig. 6. Some examples of adversarial images created with DAG attack against the detection model.

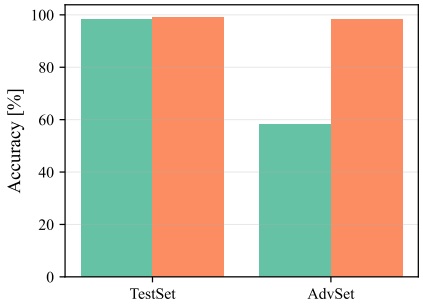

Fig. 7. Accuracy on the infrared dataset for classification with AlexNet.

3. The precision of Faster R-CNN detection model was evaluated on this adversarial dataset.

B. Second proposed object detection attack

The last experiment evaluates the power of the DAG attack against a detection model trained on infrared images. As before the detection model to be attacked is Faster R-CNN trained on the target infrared set.

We use the implementation of the Dense Generative Adversary attack for Faster R-CNN as it is described in the original paper [15], which can be found at [16]. 500 images from the dataset were chosen to create the adversarial dataset on which to evaluate the attack. The adversarial dataset obtained is then given in input into Faster R-CNN. Since it is a binary choice whether an object is a human or not, the expectations are that almost all adversarial images will be marked as empty, even if there were humans present.

V. RESULTS

Looking at the results of the first experiment, depicted in Fig. 7 and Fig. 8, two important observations can be done. Firstly, both the AlexNet trained up from the ground and the more complex pretrained ResNet have a big dropout in accuracy on the adversarial set (AdvSet) with respect to the test set (TestSet). Secondly, we can notice a significant increase in performance after adversarial training (AdvTraining) on the adversarial examples. In the AlexNet case (Fig. 7), before AdvTraining the accuracy on the TestSet and on the AdvSet were 98% and 58%. After doing the adversarial training the overall accuracy on the test set increased a little, but in particular the accuracy against adversarial examples increased significantly by a full 40%. In the ResNet case (Fig. 8), before AdvTraining, the overall accuracy on the TestSet and AdvSet were 90% and 26%, while after adversarial training even losing a bit of accuracy on the test set, the adversarial set scored an increase of accuracy of 57%.

Concerning the second experiment (on the object detection model), the baseline provided in the Detectron2 Model Zoo and Baselines repository states that for the chosen model, the Box Average Precision on COCO dataset is 40.2%.
After training on the infrared dataset, the final Box Average Precision was 63.39%. The model's predictions in fact were

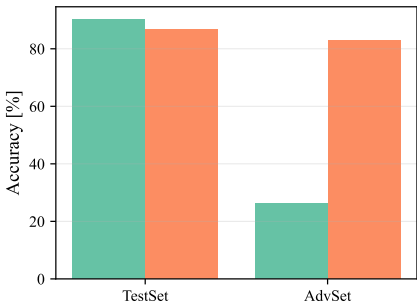

Fig. 8. Accuracy on the infrared dataset for classification with ResNet.

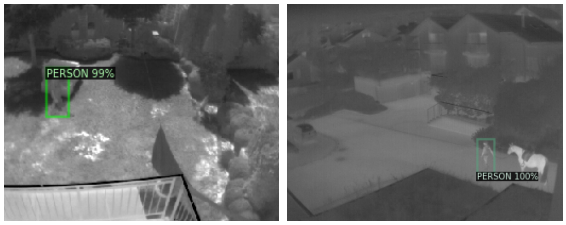

Fig. 9. Examples of Faster R-CNN output images after infrared training.

quite accurate and highly confident (we show some random examples in Fig.9).

The Faster R-CNN object detection model was then evaluated on the adversarial dataset. Surprisingly, the Average Precision on the box almost halved with respect to the benchmark on the infrared test set.

Lastly the model was tested on the adversarial dataset created with the DAG attack and results confirmed the expectations: the model was totally confused by the attack, and its high-confidence in detecting humans vanished to 0%.

Figures 10 and 11 show some example predictions of Faster R-CNN against respectively the hybrid and DAG attacks.

VI. Future Works

In further studies it would be interesting to expand the number of attacks to classification models and to see if adversarial training on Faster R-CNN makes it more robust against the used attacks. Also considering the centrality of these models in authentication systems it is important to implement more defences to all of possible attacks against models trained on infrared data.

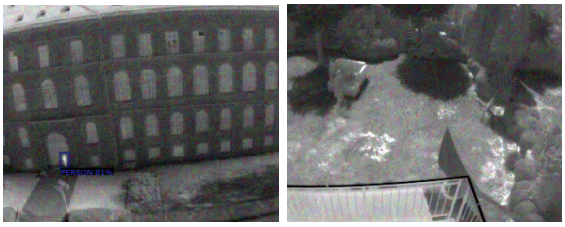

Fig. 10. Examples of output images for the first type of attack against the detection model implemented into Detectron2.

Fig. 11. Examples of output bounding boxes for the second type of attack against the detection model implemented into Detectron2.

References

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," 2014.

[2] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, pp. 14 410–14 430, 2018.

[3] E. F. J. Ring and K. Ammer, "Infrared thermal imaging in medicine," *Physiological Measurement*, vol. 33, no. 3, pp. R33–R46, feb 2012. [Online]. Available: https://doi.org/10.1088/0967-3334/33/3/r33

[4] S. D. Claudia Kuenzer, *Thermal Infrared Remote Sensing: Sensors, Methods, Applications*. Springer, 2015. [Online]. Available: https://www.springerprofessional.de/en/thermal-infrared-remote-sensing/4809380

[5] C. J. Tucker, "Red and photographic infrared linear combinations for monitoring vegetation," *Remote Sensing of Environment*, vol. 8, no. 2, pp. 127–150, 1979. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0034425779900130

[6] C. Yuan, Z. Liu, and Y. Zhang, "Fire detection using infrared images for uav-based forest fire surveillance," in *2017 International Conference on Unmanned Aircraft Systems (ICUAS)*, June 2017, pp. 567–572.

[7] B. Song, H. Choi, and H. S. Lee, "Surveillance tracking system using passive infrared motion sensors in wireless sensor network," in *2008 International Conference on Information Networking*, 2008, pp. 1–5.

[8] D. Edwards and D. B. Rawat, "Study of adversarial machine learning with infrared examples for surveillance applications," *Electronics*, vol. 9, p. 1284, 08 2020.

[9] Z. Zhou, D. Tang, X. Wang, W. Han, X. Liu, and K. Zhang, "Invisible mask: Practical attacks on face recognition with infrared," 2018.

[10] ASL-Datasets, "Icra 2014 - thermal infrared dataset," in *ASL Datasets*, 2014. [Online]. Available: https://projects.asl.ethz.ch/datasets/doku.php?id=ir:iricra2014

[11] S. Komatsu, A. Markman, A. Mahalanobis, K. Chen, and B. Javidi, "Three-dimensional integral imaging and object detection using long-wave infrared imaging," *Appl. Opt.*, vol. 56, no. 9, pp. D120–D126, Mar 2017. [Online]. Available: http://www.osapublishing.org/ao/abstract.cfm?URI=ao-56-9-D120

[12] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," 2016.

[14] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," 2017. [Online]. Available: https://arxiv.org/pdf/1612.03144.pdf

[15] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille, "Adversarial examples for semantic segmentation and object detection," 2017.

[16] yizhe ang, "Dense Adversary Generative repository," https://github.com/yizhe-ang/detectron2-1, accessed: 2021-09-30.