On the Defense Against Adversarial Examples Beyond the Visible Spectrum

Anthony Ortiz[1,†] Olac Fuentes[1,‡] Dalton Rosario[2,§] Christopher Kiekintveld[1,*]
[1]*The University of Texas at El Paso*
Computer Science Department
[†]amortizcepeda@miners.utep.edu {[‡]ofuentes, [*]cdkiekintveld}@utep.edu
[2]*U.S. Army Research Laboratory*
Image Processing Branch
[§]dalton.s.rosario.civ@mail.mil

***Abstract*—Machine learning (ML) models based on RGB images are vulnerable to adversarial attacks, representing a potential cyber threat to the user. Adversarial examples are inputs maliciously constructed to induce errors by ML systems at test time. Recently, researchers also showed that such attacks can be successfully applied at test time to ML models based on multispectral imagery, suggesting this threat is likely to extend to the hyperspectral data space as well. Military communities across the world continue to grow their investment portfolios in multispectral and hyperspectral remote sensing, while expressing their interest in machine learning based systems. This paper aims at increasing the military community's awareness of the adversarial threat and also in proposing ML training strategies and resilient solutions for state of the art artificial neural networks. Specifically, the paper introduces an adversarial detection network that explores domain specific knowledge of material response in the shortwave infrared spectrum, and a framework that jointly integrates an automatic band selection method for multispectral imagery with adversarial training and adversarial spectral rule-based detection. Experiment results show the effectiveness of the approach in an automatic semantic segmentation task using Digital Globe's WorldView-3 satellite 16-band imagery.**

***Index Terms*—Adversarial Examples, Adversarial Machine Learning, Multispectral Imagery, Defenses**

I. Introduction

Machine learning is enabling numerous innovations in many different areas and industries, including health care, transportation and logistics, among many others. New algorithms for face recognition, cancer diagnosis, and self-driving cars are just a few examples that show the progress in the field. The broad use of machine learning makes it important to understand the extent to which machine learning algorithms are subject to attacks, particularly when used in applications where physical security or safety are at risk.

Several sensitive applications, such as the screening systems in airports, military applications for mission planning, situational awareness, and surveillance, night vision systems, thermal sensors, and target identification systems, rely on sophisticated non-RGB imaging systems. Malfunctioning of these systems could have catastrophic implications. Recently, it has been shown that RGB and non-RGB image-based machine learning (ML) systems are often vulnerable to adversarial examples [3], [8], [9]. Adversarial examples are inputs maliciously constructed by adversaries to force misbehavior in the ML systems at test time. Researchers are putting efforts towards developing successful defenses against deceptive attacks in both RGB and non-RGB context [2], [8], but the problem remains largely unsolved.

In this paper, we improve on an existing defense technique to make multispectral image based ML systems for semantic segmentation robust against adversarial examples. We introduce a detection network that uses domain knowledge information of material response in the shortwave infrared (SWIR) spectrum to effectively detect adversarial examples. We also propose a framework that integrates input subset feature selection, adversarial training, and

a detector network to drastically improve the robustness of the models without sacrificing performance.

II. Background

A. Generating Adversarial Examples

The objective of adversarial learning is to find a perturbation $\boldsymbol{\xi}$ that when added to an input $\mathbf{X}$ changes the output of the model in a desired way. The attacker tries to keep $\boldsymbol{\xi}$ small enough such that when it is added to $\mathbf{X}$ to produce $\mathbf{X}^{\text{Adv}}$ = $\mathbf{X}$ + $\boldsymbol{\xi}$ the difference between $\mathbf{X}^{\text{Adv}}$ and $\mathbf{X}$ is almost imperceptible.

We denote by the function $\boldsymbol{f_\theta}$ a deep neural network with parameters $\boldsymbol{\theta}$. $\boldsymbol{f_\theta(\mathbf{X})}$ is the output of $\boldsymbol{f_\theta}$ when receiving input $\mathbf{X}$, and $\boldsymbol{y^{true}}$ is the corresponding ground-truth label. In this work, $\mathbf{X}$ is an image, $\boldsymbol{f_\theta(\mathbf{X})}$ is the conditional probability $\boldsymbol{p(y|\mathbf{X};\theta)}$ encoded as a class probability vector, and $\boldsymbol{y^{true}}$ is a one-hot encoding representation of the class. $\boldsymbol{J(f_\theta(\mathbf{X}), y^{true})}$ is the loss function. We assume that $\boldsymbol{J}$ is differentiable with respect to $\boldsymbol{\theta}$ and with respect to $\mathbf{X}$.

We tested the following attack methods:

1) Fast Gradient Sign Method (FGSM): Goodfellow et al. [3] proposed a fast single-step method for computing untargeted adversarial perturbations. This method defines an adversarial perturbation as the direction in image space that yields the greatest increase in the linearized cost function under L_∞ norm with the perturbation bounded by the parameter ϵ. This can be achieved by performing one step in the gradient sign's direction with step-width ϵ:

$$\mathbf{X}^{\text{Adv}} = \mathbf{X} + \epsilon\, sgn(\Delta_x J(f_\theta(\mathbf{X}), y^{true})) \quad (1)$$

This method is simple and computationally efficient compared to more complex methods but it usually has a lower success rate [5].

2) One-step Target Class Method (FGSM ll): Kurakin et al. [4] proposed an alternative approach to FGSM that maximizes the conditional probability $p(y^{tgt}|\mathbf{X})$ of an specific target class y^{tgt} which is unlikely to be the real class for the input image $\mathbf{X}$.

$$\mathbf{X}^{\text{Adv}} = \mathbf{X} - \epsilon\, sgn(\Delta_x J(f_\theta(\mathbf{X}), y^{tgt})) \quad (2)$$

As proposed in [4], we choose the least likely class predicted by the model as the target class y^{tgt}.

3) Iterative FGSM Method: [5], [7] This is an extension of FGSM in which FGSM is applied multiple times with a small step size:

$$\mathbf{X}_0^{Adv} = \mathbf{X},$$

$$\mathbf{X}_{i+1}^{Adv} = Clip_X, \epsilon\, \{\mathbf{X}_i^{Adv} + \alpha\, sgn(\Delta_x J(f_\theta(\mathbf{X}_i^{Adv}), y^{true}))\} \quad (3)$$

This increases the chance of fooling the original network. In this work, as in [4], we used $\alpha = 1$, which means that we changed the value of each pixel by 1 on each step. We set the number of iterations to be $min(\epsilon + 4, 1.25 * \epsilon)$.

$Clip_X, \epsilon(A)$ refers to the element-wise clipping of A, with $A_{i,j}$ clipped to the range $[\mathbf{X}_{i,j} - \epsilon, \mathbf{X}_{i,j} + \epsilon]$. This guarantees that the max l_∞-norm of the perturbation is never greater than ϵ.

4) Iterative Least-Likely Class (Iterative FGSM ll): Proposed on [5], Iterative FGSM ll is a stronger version of FGSM ll. In this case the target class is set to be the least-likely class (y^{ll}) predicted by the network to fool:

$$\mathbf{X}_0^{Adv} = \mathbf{X},$$

$$\mathbf{X}_{i+1}^{Adv} = Clip_X, \epsilon\{\mathbf{X}_i^{Adv} - \alpha\, sgn(\Delta_x J(f_\theta(\mathbf{X}_i^{Adv}), y^{ll}))\} \quad (4)$$

We used the FGSM, FGSM ll, Iterative FGSM and Iterative FGSM ll for Semantic Segmentation attacks. The attacks were generated with l_∞ norms of 2, 4, 8, 16, and 32, which corresponds to allowing increasingly more perceptible changes to the original image.

B. Multispectral Image Classification

Semantic segmentation consists of inferring labels for every pixel in an image. In the end, each pixel is labeled with the class of the enclosing object or region. The per-pixel labeling problem can be reduced to the following formulation: assign a label from the label space $L = l_1, l_2, ..., l_k$ to each element in a set of pixels $X = x_1, x_2, ..., x_N$.

Each label l represents a different class or object, e.g., building, vehicle, man-made structure, or background. This label space has k possible labels which is usually extended to $k+1$, treating l_0 as a background or void class. Usually, X is a 2D image of $W \times H = N$ pixels. However, that set can be extended to any dimensionality such as multispectral and hyperspectral images. Multispectral image classification is the task of classifying every pixel in a multispectral data cube, which is equivalent to performing semantic segmentation using a multispectral data cube as the input image. This is one of the most common uses of multispectral data so we focus on this task for our experiments.

C. Integrated Learning and Feature Selection

Ortiz et al. [8] proposed Integrated Learning and Feature Selection (ILFS) as a framework to automatically select the input features that are most useful for the learning task. Dimensionality reduction was done simultaneously with learning a model to solve the learning task. While the bands in the high dimensional image are discrete, they are densely sampled, thus they can be viewed as a continuous and differentiable space. Then, they used Stochastic Gradient Descent (SGD) to choose bands that will help a deep neural network better discriminate objects in a multispectral image. The authors showed that models trained using ILFS are robust to adversarial examples.

III. Experimental Setup

A. DSTL Satellite Imagery Dataset

The Defense Science and Technology Laboratory (DSTL) released a dataset of 1km $\times$ 1km satellite images for classification at the pixel level. There are two types of spectral imagery content provided in this dataset: 3-band images with RGB natural color and 16-band images containing spectral information captured by wider wavelength channels. This multiband imagery is taken from the Visible and Near Infrared (VNIR) (400-1040nm) and short-wave infrared (SWIR) (1195-2365nm) range collected using the DigitalGlobe's WorldView-3 satellite system. DSTL labeled 10 different classes.

B. Models

We used Tensorflow to train different models of VGG-19-based Fully Convolutional Networks (FCN-8) [6] for semantic segmentation. We trained models both with and without using ILFS for dimensionality reduction. We trained our deep network on the DSTL Satellite Image Dataset using VNIR and SWIR channels as input. 10000 randomly selected (without replacement) 224x224 patches were used for training, and 500 224x224 patches were reserved for testing. The models were trained on an NVIDIA Tesla GPU on Amazon Web Services. All the models were trained for the same number of epochs on the training set. A small batch size (4 patches) was necessary to fit the training set in memory.

C. Robustness Evaluation

The mean Intersection over Union (mean IoU) is the primary metric used for evaluating semantic segmentation. However, as the accuracy of each model varies, we adopt the relative metric used in [1] and measure adversarial robustness using the mean IoU Ratio. The mean IoU Ratio is the ratio of the network's IoU on adversarial examples to that for clean images computed over the entire dataset. A higher mean IoU Ratio implies more robustness.

IV. Non-RGB Models are Vulnerable

Ortiz et al. [8] showed that known methods to produce adversarial attacks for RGB images generalize to fool non-RGB image-based models with very little to no modifications and that it is even easier to fool this type of systems because more information can be modified.

V. Detecting Adversarial Examples

TABLE I: Detection Performance

Wetness-based Detector Network Accuracy					
Attack	$\varepsilon = 2$	$\varepsilon = 4$	$\varepsilon = 8$	$\varepsilon = 16$	$\varepsilon = 32$
FGSM	0.84	0.99	1.00	1.00	1.00
FGSM ITER	0.94	0.99	1.00	1.00	1.00
FGSM ll	0.83	0.99	1.00	1.00	1.00
FGSM ll ITER	0.95	0.99	1.00	1.00	1.00

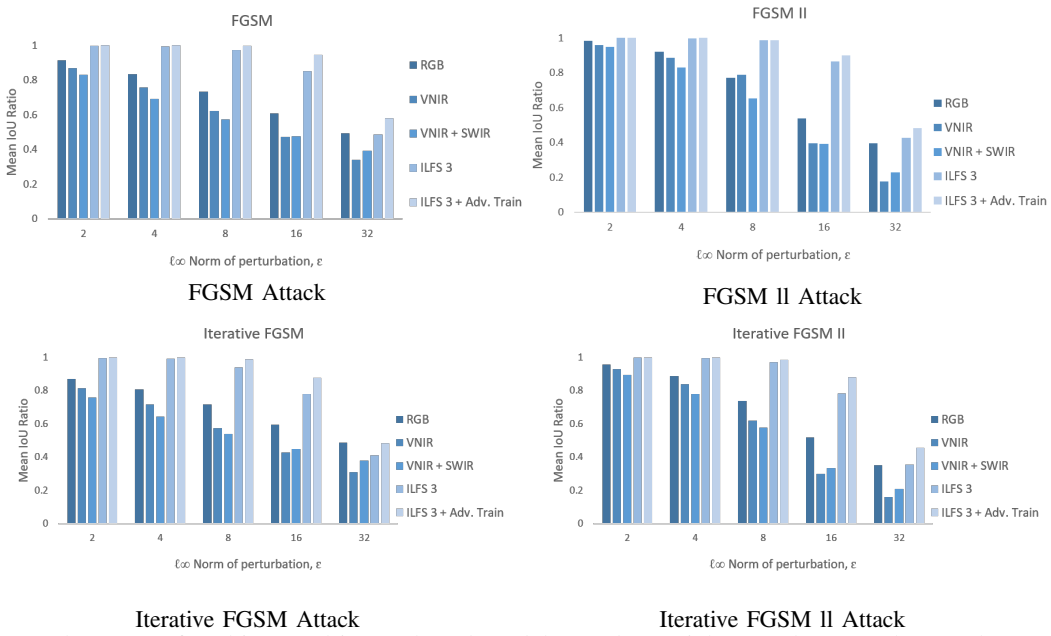


Fig. 1: Robustness of multispectral image-based models to adversarial examples. We observe that models trained on high dimensional images using ILFS along with adversarial training are more robust to adversarial examples.

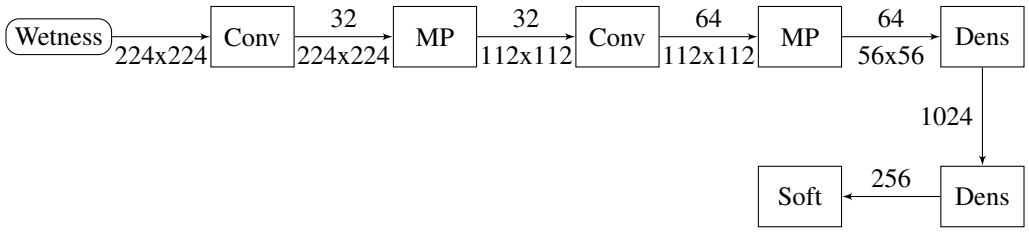


Fig. 2: Detector Network Architecture. Numbers on top of arrows denote the number of feature maps (neurons in case of dense layers) and numbers below arrows denote spatial resolutions. Conv denotes a convolutional layer, MP denotes a max pooling layer, Soft denotes softmax and Dens a fully-connected layer. Spatial resolutions are decreased by MP. All convolutional layers have 3x3 receptive fields and are followed by rectified linear units and batch normalization

Soil moisture information is one indicator of drought. Remote sensing techniques are able to record accurately the conditions of soil moisture in a large area by using a wetness index. We define a wetness index as follows:

$wetness = \frac{b_{swir2}-b_{swir4}}{b_{swir2}+b_{swir4}}$, where b_{swir2} is the image channel corresponding to the wavelength 1550-1590nm and b_{swir4} is image channel corresponding to the wavelength 1710-1750nm

The wetness index tends to be uniform among the objects in the scene so even small perturbations on different directions are easy to distinguish. The difference between the pristine image and its adversarial example from Fig. 3 is imperceptible

by humans, but our proposed detector is sensitive to that difference using as input the spectral ruled based output images shown in Fig. 3 (right side, top and bottom) instead. To exploit this fact we augment the semantic segmentation network by adding a detector subnetwork, which branches off the main network after the input layer and produces an output $p_{adv} \in [0, 1]$ which is interpreted as the probability of the input being adversarial. We train this detector network to classify the inputs as being either regular examples or examples generated by an adversary. For this, we first train the segmentation networks on the regular (non-adversarial) dataset as usual and subsequently generate adversarial examples for each data point of the train set using the FGSM method discussed previously with $\varepsilon = 8$. We thus obtain a balanced, binary classification dataset of twice the size of the original dataset consisting of the original data (label zero) and the corresponding adversarial examples (label one). From the resulting dataset we obtain the images corresponding to the wetness index of both clean and adversarial data. Then, we freeze the weights of the segmentation network and train the detector such that it minimizes the cross-entropy of p_{adv} and the labels. Fig. 2 shows the architecture used for the detector network.

The overall approach enables a machine to determine whether the input image is a pristine or adversarial example, and would only allow the former to be tasked by the classifier. Table I shows the detectability of different adversaries for the proposed detector network. Even though the detector was trained using adversarial examples from a specific attack and a specific norm of perturbation it is able to generalize to other attacks. For a perturbation greater or equal to 8, the detector networks always detect the attacks. It is important to mention that for our experiment we assume static adversary, which means that the adversary only has access to the classification network but not to the detector.

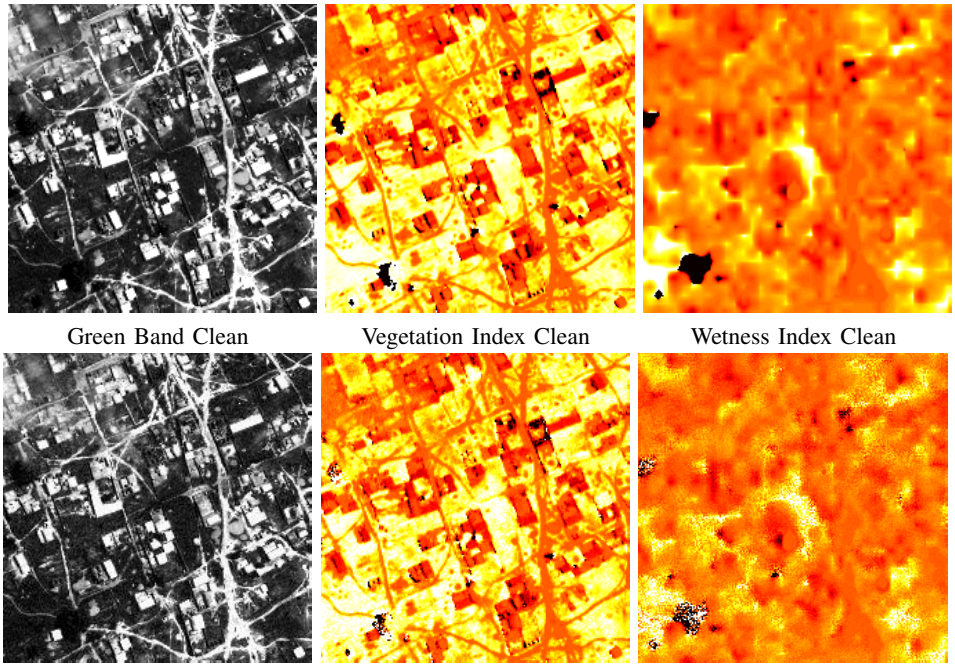
VI. Improving ILFS Robustness through Adversarial Training

Adversarial training increases robustness by augmenting training data with adversarial examples. Madry et al. [7] showed that adversarially trained models can be made robust to white box attacks (where the attacker has full knowledge about the task, model, and ML algorithm used by the defender) if the perturbation computed during training maximize the model's loss. We used the obtained adversarial examples from the previous section to augment the original DSTL training set. Then, we retrained the FCN-8 model for semantic segmentation using ILFS and the augmented dataset. This forces ILFS to choose not only the bands that will help a deep neural network to better discriminate objects in a multispectral image, but also to choose bands that are less sensitive to adversarial perturbations. Figure 1 shows the mean IoU ratio as a measure of the robustness of the trained models to adversarial examples obtained in a white box setting with different l_∞-norm of perturbation (2, 4, 8, 16, 32). From Figure 1 we can see that multispectral image-based models trained using ILFS and adversarial training are more robust to adversarial examples.

As a final framework for semantic segmentation of multispectral images, we propose to combine the spectral-rule based adversarial detection network with a network trained using ILFS and adversarial training. Figure 1 shows that most of the mistakes done by ILFS occur when the perturbation is bigger than 8. The adversarial detector network always detects adversaries with a perturbation bigger than 8. Because of that, both defense mechanisms complement each other producing a very robust model.

VII. Conclusions

We introduced a network that detects adversarial examples from images generated by applying a knowledge-based spectral filter in the SWIR region. The network achieves an accuracy above 83% in the detection of adversarial examples generated using for state-of-the-art algorithms. We also showed that an existing multispectral based ILFS defense could be improved by simultaneously employing adversarial training and automatic spectral band selection. Finally, we proposed a framework that integrates the SWIR rule based adversarial detection network, ILFS, and adversarial training to achieve a significantly improved ML model resilience, as



Green Band Clean | Vegetation Index Clean | Wetness Index Clean

Green Band Adversarial | Vegetation Index Adversarial | Wetness Index Adversarial

Fig. 3: Top images are clean. Bottom Images are obtained from an adversarial example. Vegetation index was obtained as follows: $NDVI = \frac{b_{vnir7} - b_{vnir5}}{b_{vnir7} + b_{vnir5}}$, where b_{vnir7} is the image channel corresponding to the wavelength 770-895nm and b_{vnir5} is image channel corresponding to the wavelength 630-690nm.

a promising approach for the military community. For follow up, we plan to extend this research to explore spectral material phenomena in the thermal longwave infrared (LWIR: 7.5-14.0 micron) spectrum for daytime and nighttime machine learning resilience to adversarial attacks.

ACKNOWLEDGEMENT

This work was supported by the Army Research Office under award W911NF-17-1-0370.We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPUs used for this research.

REFERENCES

[1] A. Arnab, O. Miksik, and P. H. Torr. On the robustness of semantic segmentation models to adversarial attacks. *arXiv preprint arXiv:1711.09856*, 2017.

[2] E. D. Cubuk, B. Zoph, S. S. Schoenholz, and Q. V. Le. Intriguing properties of adversarial examples. *arXiv preprint arXiv:1711.02846*, 2017.

[3] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *International Conference on Learning Representations (ICLR)*, 1050:20, 2015.

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

[5] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *International Conference on Learning Representations (ICLR)*, 2017.

[6] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[7] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *stat*, 1050:19, 2017.

[8] A. Ortiz, A. Granados, O. Fuentes, C. Kiekintveld, D. Rosario, and Z. Bell. Integrated learning and feature selection for deep neural networks in multispectral images. *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018.

[9] N. Papernot, P. McDaniel, A. Sinha, and M. Wellman. Towards the science of security and privacy in machine learning. *arXiv preprint arXiv:1611.03814*, 2016.