Integrated Learning and Feature Selection for Deep Neural Networks in Multispectral Images

Anthony Ortiz[1] Alonso Granados[1] Olac Fuentes[1]
Christopher Kiekintveld[1] Dalton Rosario[2] Zachary Bell[1]
[1]University of Texas at El Paso [2]U.S. Army Research Laboratory (ARL)
{amortizcepeda, agranados11, zjbell}@miners.utep.edu {ofuentes, cdkiekintveld}@utep.edu
dalton.s.rosario.civ@mail.mil

Abstract

The curse of dimensionality is a well-known phenomenon that arises when applying machine learning algorithms to highly-dimensional data; it degrades performance as a function of increasing dimension. Due to the high data dimensionality of multispectral and hyperspectral imagery, classifiers trained on limited samples with many spectral bands tend to overfit, leading to weak generalization capability. In this work, we propose an end-to-end framework to effectively integrate input feature selection into the training procedure of a deep neural network for dimensionality reduction. We show that Integrated Learning and Feature Selection (ILFS) significantly improves performance on neural networks for multispectral imagery applications. We also evaluate the proposed methodology as a potential defense against adversarial examples, which are malicious inputs carefully designed to fool a machine learning system. Our experimental results show that methods for generating adversarial examples designed for RGB space are also effective for multispectral imagery and that ILFS significantly mitigates their effect.

1. Introduction

The remote sensing community has started to adopt deep learning and apply it to multispectral and hyperspectral imagery [13, 32]. However, the very high dimensionality of these images and the limited size of available data sets for training limit the exploitation of this methodology. Deep neural networks trained with many spectral bands tend to overfit, which leads to weak generalization capability even when sufficient training data is available [27]. In addition, this high dimensionality can make the models highly susceptible to adversarially constructed inputs.

A common approach to reduce high dimensionality is to transform the data into a lower dimensional space using methods like Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA) [17, 19]. However, these methods usually change the meaning of the original data as the features in the low-dimensional space are linear combinations of the original bands. A second approach to dimensionality reduction for hyperspectral and multispectral images is band selection, which selects a subset of the original bands. Band selection techniques can be divided into two categories: supervised and unsupervised [7]. Supervised methods aim to preserve the object information related to a target function, which is known a priori [7, 43], while unsupervised methods do not assume any object information [2, 16]. All previous band selection methods follow a two-step procedure where an independent method is used initially to select a subset of bands and then a learning algorithm is run using these bands as input.

We propose a new approach for dimensionality reduction that directly integrates supervised feature selection with a deep learning classifier. We call this method Integrated Learning and Feature Selection (ILFS). ILFS allows feature selection to be optimized by the learning algorithm, and because information from bands that are not ultimately selected is available during the learning process, it can improve performance. We demonstrate that integrating feature selection into and end-to-end deep learning algorithm greatly improves performance in multispectral image classification. We also show that it is an effective defense against adversarial examples.

It has been shown recently that machine learning models based on RGB images are often vulnerable to adversarial examples [22, 25, 33, 34, 37]. Adversarial examples are inputs maliciously constructed to induce errors by machine learning models at test time. This represents a new attack vector against systems that rely on machine learning models for critical functions (e.g., facial recognition to establish identity, among many others). Researchers are working towards developing defenses against attacks on RGB-based

classifiers and have proposed numerous strategies to train models that are more robust to adversarial examples or to detect adversarial examples [6, 15, 26, 35, 29]. However, they have frequently been found to be vulnerable to other types of attacks, or come at the cost of decreased performance on clean inputs [3, 8, 9, 10, 23, 29].

Many highly sensitive applications of remote sensing and image analysis rely on more sophisticated imaging technologies that go beyond RGB. Such applications include screening systems in airport security, military applications of satellite imagery for mission planning, situational awareness, surveillance, night vision systems, thermal sensors, and target identification systems [18, 21, 31, 36, 39, 40, 41, 44, 46]. These systems are especially attractive targets for highly skilled and motivated attackers, and the consequences of adversarial attacks on learning algorithms in these domains could be catastrophic. However, we are not aware of any previous work that focuses on adversarial examples for machine learning with non-RGB images.

We present the first rigorous study of the robustness of non-RGB image-based systems (VNIR, SWIR, Panchromatic) against adversarial examples for the task of semantic segmentation. We show that it is not only feasible to generate deceptive examples for machine learning models based on non-RGB images, it is easier in a sense than attacking models for RGB images. However, we show that performing band selection using ILFS can substantially improve the robustness of machine learning models against these adversarial examples. ILFS limits the attack surface by reducing the number of features that can be modified by an attacker to induce errors, and does so in an unpredictable way. ILFS also forces the model to perform well when there is uncertainty in the input space because the input changes throughout the learning process. While we demonstrate ILFS here for band selection, this technique could be adapted to any learning task with a continuous input space.

We summarize our main contributions as follows:

- We propose Integrated Learning and Feature Selection (ILFS) as a generic framework for supervised dimensionality reduction. We demonstrate ILFS is effective for dimensionality reduction of multispectral and hyperspectral imagery, and significantly improves performance on the semantic segmentation task for high dimensional imagery.
- We present the first study of constructing adversarial examples for non-RGB imagery, and show that non-RGB machine learning models are vulnerable to adversarial examples.
- We show that Integrated Learning and Feature Selection (ILFS) is an effective defense to make multispectral image-based models more robust against adversarial examples.

2. Semantic Segmentation and Multispectral Image Classification

Semantic segmentation makes dense predictions, 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: find a way to 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 volumetric data or multispectral and hyperspectral images. Multispectral image classification is the task of classifying every pixel in a multispectral data cube, which is similar 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.

3. Integrated Learning and Feature Selection

We propose ILFS as a framework to automatically select the input features that are most useful for the learning task. Dimensionality reduction is done simultaneously with learning a model to solve the learning task. In the case of our semantic segmentation task, this corresponds to choosing bands that will help a deep neural network better discriminate objects in a multispectral image.

Let us consider I as the input space with n features and let Z be the vector encoding the selected features $\{z_1, z_2, \dots z_{k-1}, z_k\}$ (e.g., the bands in a multispectral or hyperspectral image). We define X to be the image that results from selecting features Z from I, and J to be the cost function of the network. To discover and select features that can discriminate objects more effectively, we include the input feature selection in the learning process. To achieve this, we compute the gradient of the loss with respect to the selected features using the chain rule:

$$\frac{\partial J}{\partial Z} = \frac{\partial J}{\partial X}\frac{\partial X}{\partial Z} \tag{1}$$

3.1. Obtaining the Derivatives

To compute $\frac{\partial J}{\partial X}$, it is only necessary to backpropagate the loss from the last layer up to the input image. To obtain $\frac{\partial X}{\partial Z}$, we compute $\frac{\partial X}{\partial z_i}$ for each of the features in Z. 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. We compute the values of frac-

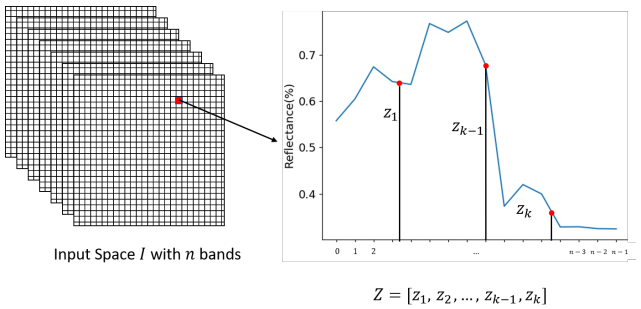


Figure 1: $H(I, Z)$ for a pixel by selecting k features

tional bands using simple linear interpolation, which leads to piecewise constant derivatives (see Figure 1). Thus, if we define $H(I, z)$ to be the resulting image from selecting feature z from I, equation 2 shows how a change in a particular feature z_i produces a change in X.

$$\frac{\partial X}{\partial z_i} = H(I, \lfloor z_i \rfloor + 1) - H(I, \lfloor z_i \rfloor) \quad (2)$$

Finally, we define $\frac{\partial X}{\partial Z}$, as the vector with the values $\left[\frac{\partial X}{\partial z_1}, \frac{\partial X}{\partial z_2}, \cdots \frac{\partial X}{\partial z_{k-1}}, \frac{\partial X}{\partial z_k}\right]$

3.2. ILFS Implementation Details

Our implementation of ILFS initially defines Z as a random vector with the bands selected. We use this vector to obtain the image X from the input image I that has all bands. We update the bands 10 times per epoch using gradient descent and an adaptive learning rate. This is necessary to prevent premature convergence in the model. We use an initial learning rate of 0.2 and decrease it exponentially after every epoch using exponential decay. We update X every time Z is updated.

4. Experimental Setup

4.1. DSTL Satellite Imagery Dataset

The Defense Science and Technology Laboratory (DSTL) released a dataset of 1km $\times$ 1km satellite images for detection and classification of the types of objects found in these regions 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 multi-band imagery is taken from the Visible and Near Infrared (VNIR) (400-1040nm) and short-wave infrared (SWIR) (1195-2365nm) range collected using the DigitalGlobes WorldView-3 satellite system. DSTL labeled 10 different classes:

1. **Buildings**: large building, residential, non-residential, fuel storage facility, fortified building.
2. **Misc**: manmade structures.
3. **Road**
4. **Track**: poor/dirt/cart track, footpath, trail.
5. **Trees**: woodland, hedgerows, groups of trees, standalone trees.
6. **Crops**: contour ploughing/cropland, grain crops (wheat, corn), row crops (potatoes, turnips).
7. **Waterway**
8. **Standing Water**
9. **Large, Vehicle**: large vehicle (e.g. lorry, truck, bus), logistics vehicle.
10. **Small Vehicle**: small vehicle (car, van), motorbike.

4.2. Models

In semantic segmentation, we want to assign each pixel in the input image to an object class. Most popular approaches to do semantic segmentation are based on Fully Convolutional Networks (FCN) [28]. FCN are a type of Convolutional Neural Network architecture for dense predictions that do not use any fully connected layers. This allows segmentation maps to be generated for large images. Almost all subsequent state-of-the-art approaches for semantic segmentation have adopted this paradigm [4, 5, 12].

We used Tensorflow to train different models of VGG-19-based FCN-8 for semantic segmentation [28, 38]. We trained models both with and without using ILFS for dimensionality reduction. We trained our deep network on the DSTL Satellite Image Dataset using either RGB, VNIR, SWIR, or 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.

4.3. ILFS Evaluation

We report performance results using mean intersection over the union (mean IoU), a standard metric for common semantic segmentation and scene parsing evaluations.

$$meanIoU = (1/n_{cls}) \sum_i n_{ii}/(t_i + \sum_j n_{ji} - n_{ii}) \quad (3)$$

where n_{ij} is the number of pixels of class i predicted to belong to class j, there are n_{cls}different classes, and $t_i = \sum_j n_{ji}$ is the total number of pixels of class i.

5. Results

5.1. ILFS Significantly Improves Performance

Figure 2 presents the results of training models for semantic segmentation using different inputs, including ILFS

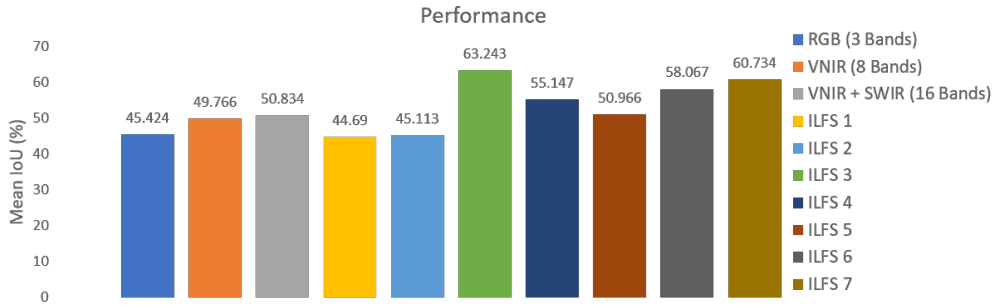


Figure 2: Performance Evaluation for Semantic Segmentation of Hyperspectral Images. All of models were trained on the same deep neural network architecture (VGG19-based FCN-8) for the same number of epochs. For RGB we trained using only the three RGB bands as input. For VNIR, we trained with all of the bands from the visible and near-infrared (VNIR) region of the spectrum. For VNIR + SWIR all of the available bands were used as input while training. ILFS 1...7 show the results when ILFS is used to pick 1, 2, 3, 4, 5, 6, and 7 input features, respectively.

used to select different numbers of input features. Performance is measured using the mean IoU. All of the models used the same network architecture (VGG19-based FCN-8) and the same number of training epochs. The results show that ILFS improves performance by up to 12.409% over the best model without ILFS. All ILFS models with more than 2 features beat all of the non-ILFS models. Hyper-parameters were fine-tuned for ILFS 3 and then used for the other experiments. Improvement in performance could be obtained by fine-tuning hyper-parameters for every model. The intuition for the improved accuracy is that ILFS allows the network to find a combination of bands (including interpolated bands) that allow for better discrimination of the objects in the scene while the smaller feature space prevents overfitting of the training data.

Another advantage of ILFS with 3 features is the feasibility of doing transfer learning. Training a deep neural network from scratch may not be feasible for various reasons: a dataset of sufficient size may not be available, or reaching convergence can take too long, or the memory requirement may be too high for the available hardware. Even if training a network is feasible, it is often helpful to start with pre-trained weights instead of randomly initialized weights. Here we can use a model previously trained on RGB for a similar task to initialize the weights and selected features and then continue the training using ILFS 3. Yosinski et al. showed that transferring features even from distant tasks can be better than using random initialization, taking into account that the transferability of features decreases as the difference between the pre-trained task and the target one increases [45].

5.2. ILFS Visualization

For ILFS 3 it is possible to display a false color image of the selected bands. This allows the user to visualize what the network has learned and determine if it is a good set of discriminative channels or bands. Figure 3 shows the true color image of one portion of the data cube and the corresponding false-color image of the features obtained from ILFS 3. This output is very similar to what is known as spectral indices in remote sensing. We can visually discriminate classes (e.g., vegetation is bright white), which may be useful for human analysts as well.

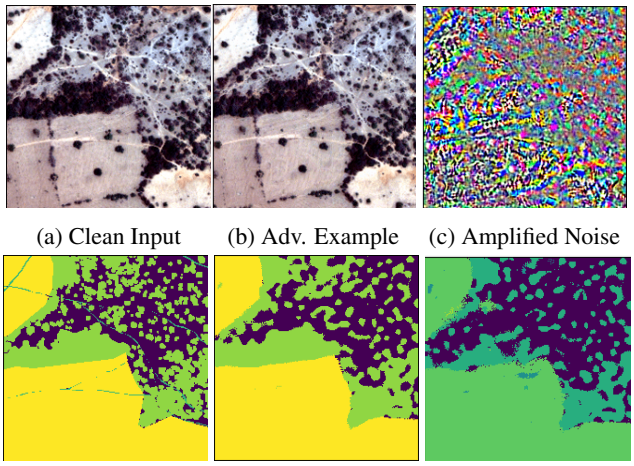
(a) True Color (b) False Color ILFS 3

Figure 3: ILFS 3 Visualization. We see similar results to what is known as spectral indices in the remote sensing community, and can visually discriminate classes.

6. Adversarial Examples Beyond RGB

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

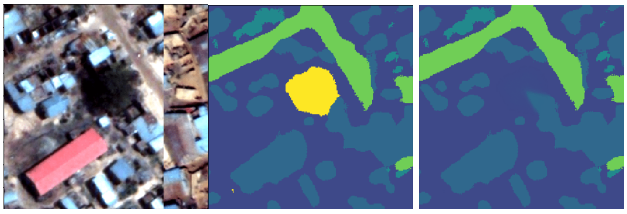
We denote by the function f_θ a deep neural network with parameters θ. $f_\theta(\mathbf{X})$ is the output of f_θ, and y^{true} is the corresponding ground-truth label. In this work, x is an image, $f_\theta(\mathbf{X})$ is the conditional probability $p(y|\mathbf{X};\theta)$ encoded as a class probability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



(a) Clean Input (b) Adv. Example (c) Amplified Noise

(d) Ground-Truth (e) Pred. Clean (f) Pred. Adversarial

Figure 4: An adversarial example generated with l_∞-norm of 4. (a) RGB of the original image (b) RGB representation of the adversarial examples obtained using the Iterative FGSM ll method (c) the amplified noise added to the original image (d) the ground-truth image (e) the model prediction when the original image is the input (f) the model prediction when the adversarial example is the input.



(a) True Color Input (b) Pred. Clean (c) Pred. Adversarial

Figure 5: Dynamic Adversarial Perturbation for HSI Semantic Segmentation attack.

6.1. Methods to Generate Adversarial Examples

We tested the following attack methods:

Fast Gradient Sign Method (FGSM): Goodfellow et al. [22] 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:

$$\mathbf{X}^{\text{Adv}} = \mathbf{X} + \epsilon\, sign(\Delta_x J(f_\theta(\mathbf{X}), y^{true})) \quad (4)$$

This method is simple and computationally efficient compared to more complex methods but it usually has a lower success rate [25].

One-step Target Class Method (FGSM ll): Kurakin et al. [24] proposed an alternative approach to FGSM that maximizes the conditional probability $p(y^{target}|\mathbf{X})$ of an specific target class y^{target} which is unlikely to be the real class for the input image $\mathbf{X}$.

$$\mathbf{X}^{\text{Adv}} = \mathbf{X} - \epsilon\, sign(\Delta_x J(f_\theta(\mathbf{X}), y^{target})) \quad (5)$$

As proposed in [24], we choose the least likely class predicted by the model as the target class y^{target}.

Basic Iterative Method (Iterative FGSM): [25, 29] This is an extension of FGSM in which FGSM is applied multiple times with a small step size:

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

$$\mathbf{X}_{i+1}^{Adv} = Clip_X, \epsilon\{\mathbf{X}_i^{Adv} + \alpha\, sign(\Delta_x J(f_\theta(\mathbf{X}_i^{Adv}), y^{true}))\} \quad (6)$$

This increases the chance of fooling the original network. In this work, as in [24], 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.

Iterative Least-Likely Class (Iterative FGSM ll) [25] 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:

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

$$\mathbf{X}_{i+1}^{Adv} = Clip_X, \epsilon\{\mathbf{X}_i^{Adv} - \alpha\, sign(\Delta_x J(f_\theta(\mathbf{X}_i^{Adv}), y^{ll}))\} \quad (7)$$

we used $\alpha = 1$ and the number of iterations was set to $min(\epsilon + 4, 1.25 * \epsilon)$.

These attacks were all originally proposed in the context of RGB image classification, but they have been adapted to semantic segmentation [14, 20, 30, 42], object detection [42], and other tasks.

Dynamic Adversarial Perturbations for Semantic Segmentation: For semantic segmentation, the loss function is a sum over the spatial dimensions of the ground-truth.

$$J_s(f_\theta(\mathbf{X}), y) = 1/nmPix \sum_{(i,j)\in X} J_{cls}(f_\theta(\mathbf{X})_{ij}, y_{ij}) \quad (8)$$

Metzen et al. [30] describes an adversarial example for semantic segmentation as an input x_{adv} for f_θ such that

$J_s(f_\theta(\mathbf{X}), y^{tgt})$ is minimal without making perceptible changes to the input. In the context of multispectral and hyperspectral images in addition to keeping the spatial information almost identical to the input, the spectral signature of every pixel should be preserved. Otherwise, experts could identify the perturbations by just looking at the spectral information. Real world scenarios in remote sensing may consist of an adversary trying to hide certain kind of object. We assume that the adversary has access to the model f_θ, so he can use $y^{pred} = f_\theta(\mathbf{X})$ as an initial step, and he would like to keep y^{tgt} as similar as possible to y^{pred} to avoid attracting the attention of humans monitoring the system. To accomplish this Metzen et al. proposed assigning to the target class the predicted output for all the pixels in the background (X_{bg}) (the ones you are not looking to hide) and filling the gaps of the objects trying to hide (X_o) by interpolating pixels in the background using a nearest-neighbor heuristic. We follow the same idea for y^{tgt} in this work.

Given y^{tgt}, adversarial examples to hide objects while making the spatial and spectral changes imperceptible can be obtained using the following formulation:

$$J_s(f_\theta(X), y) = 1/nmPix\{w \sum_{(i,j)\in X_o} J_{cls}(f_\theta(X)_{ij}, y_{ij}^{tgt}) + (1-w) \sum_{(i,j)\in X_{bg}} J_{cls}(f_\theta(X)_{ij}, y_{ij}^{tgt}\} \tag{9}$$

$$\mathbf{X}_0^{Adv} = \mathbf{X},$$
$$\mathbf{X}_{i+1}^{Adv} = Clip_{X,\epsilon}\{\mathbf{X}_i^{Adv} - \alpha\, sign(\Delta_x Js(f_\theta(\mathbf{X}_i^{Adv}), y^{tgt}))\} \tag{10}$$

6.2. Attacks Used

We used the FGSM, FGSM ll, Iterative FGSM, Iterative FGSM ll, and Dynamic Adversarial perturbations 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.

6.3. 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.

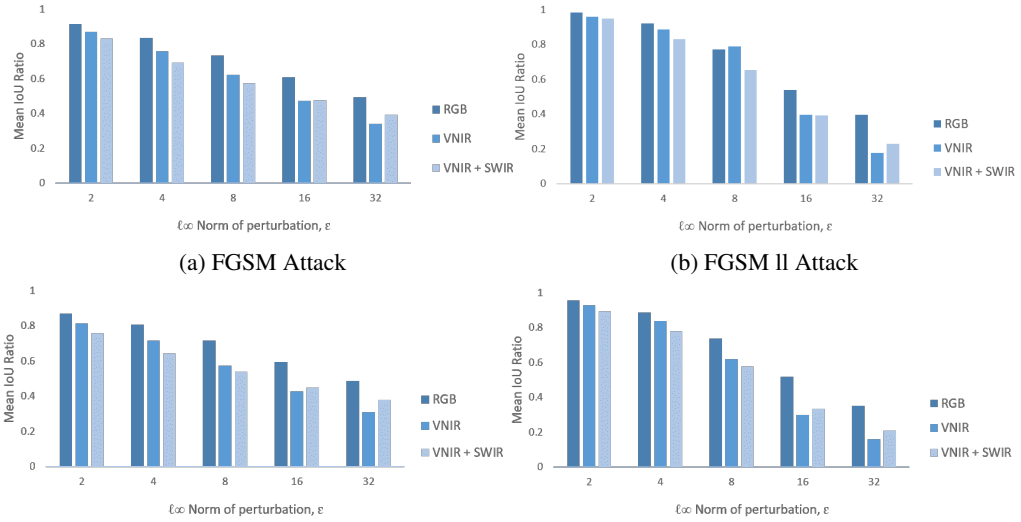
6.4. Non-RGB Image-Based Models are Vulnerable to Adversarial Examples

Figure 4 shows an example of an attack on a multispectral model using the Iterative FGSM ll with an l_∞-norm of perturbation of 4. To visualize the results we show the true color composition for the multispectral clean and adversarial input. The difference between the clean and the adversarial input is visually imperceptible, but the predictions of the model are totally different. FGSM, Iterative FGSM, FGSM ll, and Iterative FGSM ll attacks try to cause the model to make as many mistakes as possible. The main issue with these attacks is that in real life scenarios totally disassociated with the real classes will make the attacks obvious. Attackers will likely prefer attacks like the one shown in Figure 5. This attack can be used to hide specific classes and/or objects from the scene while giving as output a prediction that is as close as possible to the real prediction. Figure 5 shows how we successfully hid a tree for the adversarial prediction.

Figure 6 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 6 we can see that multispectral image-based models are vulnerable to adversarial examples. In fact, it is even easier to fool those models in a white box setting, as they produce lower mean IoU ratio than RGB models for the same amount of perturbation. The intuition behind this result is that with high dimensional images an attacker has more information to manipulate.

6.5. Non-RGB Adversarial Examples in the Physical World

Adversarial attacks have proven to be successful in the physical world as well [11, 37]. Adversarial examples in the physical world are normally accomplished by printing the color image of the adversarial examples. To test this, we generated adversarial examples against an RGB image-based semantic segmentation model and used those adversarial examples to modify the RGB part of the input images sent to the model trained on both Visible Near Infrared (VNIR) and Short Wave Infrared (SWIR) images. This is an attempt to study attacks in settings where the attacker is constrained in the information that can be manipulated. In the physical world, modifying RGB is trivial, but modifications in other regions of the spectrum like near infrared are more difficult because other physical conditions like the temperature of the objects need to be manipulated as well. Figure 7 provided evidence that this type of attack will not be successful when attacking models that include more spectral information.



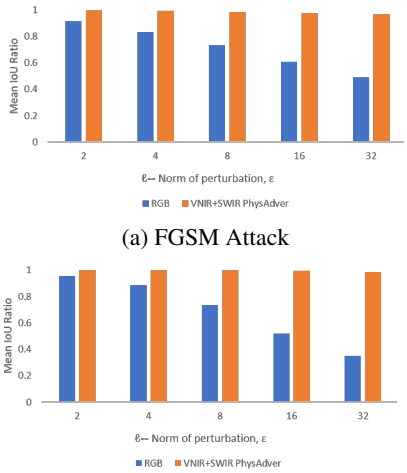
(a) FGSM Attack

(b) FGSM ll Attack

(c) Iterative FGSM Attack

(d) Iterative FGSM ll Attack

Figure 6: Robustness of multispectral image-based models to adversarial examples. We observe that models trained on high dimensional images are even more vulnerable to adversarial examples than RGB image-based models for white box settings for all four tested attacks.



(a) FGSM Attack

(b) Iterative FGSM ll Attack

Figure 7: Attacks in the Physical World. For these set of attacks, the attacker can only manipulate the visible region of the spectrum.

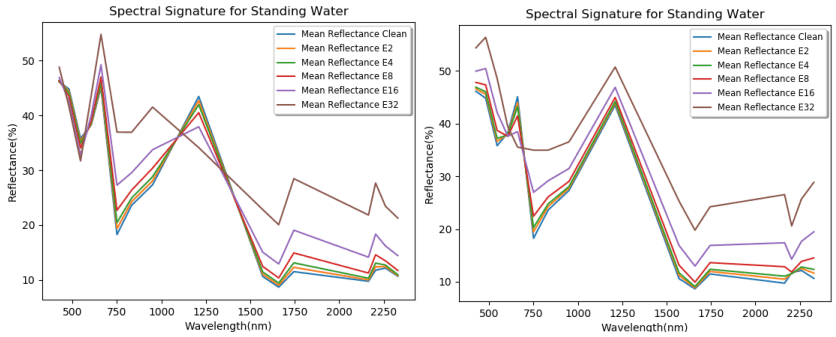
6.6. Spectral Signature of Adversarial Examples

It is possible to obtain the spectral signature of the different materials from high dimensional imagery. Figure 8 shows the mean reflectance for the pixels belonging to the class "standing water" for the clean input images and adversarial examples crafted using different l_∞-norm of perturbation for two different attacks. Figure 8 offers an intuition for what the attack is doing on the input image. Larger l_∞-norm perturbations produce more drastic changes in the spectral signature of the class. This could be exploited to actually detect these adversarial perturbations.

6.7. ILFS as a Defense to Adversarial Examples

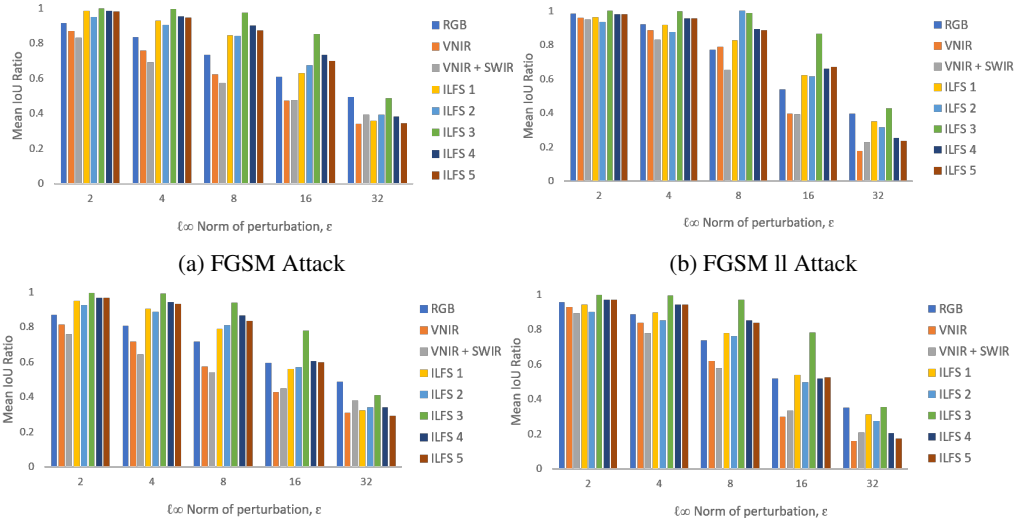
To test the robustness of ILFS to adversarial examples, we attacked ILFS models obtained selecting different amount of features with four attack methods and different l_∞-norm of perturbation. We use the mean IoU Ratio as a metric to compare robustness between ILFS models and models trained in the traditional way on RGB, VNIR, VNIR + SWIR regions of the spectrum. Figure 9 shows that, in general, ILFS models not only achieve better performance on clean inputs but are also more robust to adversarial examples as their mean IoU ratio is consistently higher. This is particularly the case for all l_∞-norm of perturbation smaller than 32. When the l_∞-norm of perturbation is 32 the margin between the most and least robust model is smaller as none of them perform well. For l_∞-norm of 32 the perturbations are visually obvious and the spectral signature of the different classes is drastically modified (See Figure 8) which will not represent acceptable adversarial examples. Moreover, as expected, the Iterative FGSM ll attack is more powerful at fooling networks than single-step FGSM for non-RGB image-based models.

We trained a model using the output bands from ILFS 3 as a fixed input and compare it's robustness with the model trained using Integrated Learning and Feature Selection. Figure 10 shows that the model trained on fixed inputs is less robust. This supports our hypothesis that as ILFS keeps changing the input space during training, ILFS forces the network to perform well even when the input is maliciously modified.



(a) FGSM (b) Iterative FGSM ll

Figure 8: Spectral Signature of Adversarial Examples. Larger l_∞-norm produces more drastic changes to the spectral signature of the class.



(a) FGSM Attack (b) FGSM ll Attack

(c) Iterative FGSM Attack (d) Iterative FGSM ll Attack

Figure 9: ILFS as a Defense to Adversarial Examples

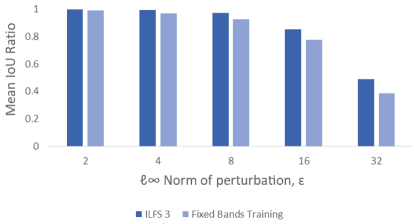


Figure 10: Robustness of ILFS 3 vs training on fixed input.

7. Conclusions

In this study, we have introduced Integrated Learning and Feature Selection (ILFS) as a framework for dimensionality reduction of high dimensional imagery through supervised feature subset selection using gradient descent on the input space. ILFS not only reduces data dimensionality but also improves performance on deep neural networks for multispectral imagery applications. ILFS is general enough to be extensible to any machine learning problem with continuous input space.

We have shown what, to the best of our knowledge, is the first rigorous evaluation of the robustness of non-RGB image-based machine learning models to adversarial attacks. We 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. In fact, it is even easier to fool this type of systems because more information can be modified. Adversarial examples in the physical world are more difficult to execute on non-RGB image-based models because in those settings the attacker will be required to manipulate not only the color but other properties (i.e. temperature) of the objects in the scene.

Finally, we showed that applying IFLS increases robustness to adversarial examples in the high dimensional semantic segmentation problem, considering four state-of-the-art attack algorithms.

Acknowledgement

This work was supported by the Army Research Office under award W911NF-17-1-0370.

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. 6

[2] M. G. Asl, M. R. Mobasheri, and B. Mojaradi. Unsupervised feature selection using geometrical measures in prototype space for hyperspectral imagery. *IEEE transactions on geoscience and remote sensing*, 52(7):3774–3787, 2014. 1

[3] A. Athalye, N. Carlini, and D. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018. 2

[4] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481, 2017. 3

[5] M. Bai and R. Urtasun. Deep watershed transform for instance segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2858–2866. IEEE, 2017. 3

[6] J. Bradshaw, A. G. d. G. Matthews, and Z. Ghahramani. Adversarial examples, uncertainty, and transfer testing robustness in gaussian process hybrid deep networks. *arXiv preprint arXiv:1707.02476*, 2017. 2

[7] X. Cao, T. Xiong, and L. Jiao. Supervised band selection using local spatial information for hyperspectral image. *IEEE Geoscience and Remote Sensing Letters*, 13(3):329–333, 2016. 1

[8] N. Carlini and D. Wagner. Defensive distillation is not robust to adversarial examples. *arXiv preprint arXiv:1607.04311*, 2016. 2

[9] N. Carlini and D. Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 3–14. ACM, 2017. 2

[10] N. Carlini and D. Wagner. Magnet and" efficient defenses against adversarial attacks" are not robust to adversarial examples. *arXiv preprint arXiv:1711.08478*, 2017. 2

[11] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *Security and Privacy (SP), 2017 IEEE Symposium on*, pages 39–57. IEEE, 2017. 6

[12] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 3

[13] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu. Deep learning-based classification of hyperspectral data. *IEEE Journal of Selected topics in applied earth observations and remote sensing*, 7(6):2094–2107, 2014. 1

[14] M. Cisse, Y. Adi, N. Neverova, and J. Keshet. Houdini: Fooling deep structured prediction models. *arXiv preprint arXiv:1707.05373*, 2017. 5

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

[16] Q. Du and H. Yang. Similarity-based unsupervised band selection for hyperspectral image analysis. *IEEE Geoscience and Remote Sensing Letters*, 5(4):564–568, 2008. 1

[17] Q. Du and N. H. Younan. Dimensionality reduction and linear discriminant analysis for hyperspectral image classification. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pages 392–399. Springer, 2008. 1

[18] M. Ettenberg. A little night vision-ingaas shortwave infrared emerges as key complement to ir for military imaging. *Advanced Imaging-Fort Atkinson*, 20(3):29–33, 2005. 2

[19] M. D. Farrell and R. M. Mersereau. On the impact of pca dimension reduction for hyperspectral detection of difficult targets. *IEEE Geoscience and Remote Sensing Letters*, 2(2):192–195, 2005. 1

[20] V. Fischer, M. C. Kumar, J. H. Metzen, and T. Brox. Adversarial examples for semantic image segmentation. *International Conference on Learning Representations (ICLR) Workshop*, 2017. 5

[21] M. G. Glaholt and G. Sim. Gaze-contingent center-surround fusion of infrared images to facilitate visual search for human targets. *Journal of Imaging Science and Technology*, 61(1):10401–1, 2017. 2

[22] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *International Conference on Learning Representations (ICLR)*, 1050:20, 2015. 1, 5

[23] S. Gu and L. Rigazio. Towards deep neural network architectures robust to adversarial examples. *International Conference on Learning Representations (ICLR) Workshop*, 2015. 2

[24] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016. 5

[25] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *International Conference on Learning Representations (ICLR)*, 2017. 1, 5

[26] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6393–6395, 2017. 2

[27] F. Li, L. Xu, P. Siva, A. Wong, and D. A. Clausi. Hyperspectral image classification with limited labeled training samples using enhanced ensemble learning and conditional random fields. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(6):2427–2438, 2015. 1

[28] 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. 3

[29] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *stat*, 1050:19, 2017. 2, 5

[30] J. H. Metzen, M. C. Kumar, T. Brox, and V. Fischer. Universal adversarial perturbations against semantic image segmentation. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017. 5

[31] P. Pallister, T. DSouza, C. Black, N. Hearns, and J. C. Smith. Explosive detection strategies for security screening at airports. In *Molecular Technologies for Detection of Chemical and Biological Agents*, pages 243–251. Springer, 2017. 2

[32] B. Pan, Z. Shi, and X. Xu. Mugnet: Deep learning for hyperspectral image classification using limited samples. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2017. 1

[33] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. The limitations of deep learning in adversarial settings. In *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*, pages 372–387. IEEE, 2016. 1

[34] 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. 1

[35] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *Security and Privacy (SP), 2016 IEEE Symposium on*, pages 582–597. IEEE, 2016. 2

[36] D. A. Robertson, D. G. Macfarlane, R. I. Hunter, S. L. Cassidy, N. Llombart, E. Gandini, T. Bryllert, M. Ferndahl, H. Lindström, J. Tenhunen, et al. High resolution, wide field of view, real time 340ghz 3d imaging radar for security screening. In *Passive and Active Millimeter-Wave Imaging XX*, volume 10189, page 101890C. International Society for Optics and Photonics, 2017. 2

[37] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter. 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*, pages 1528–1540. ACM, 2016. 1, 6

[38] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*, 2015. 3

[39] D. Stein, J. Schoonmaker, and E. Coolbaugh. Hyperspectral imaging for intelligence, surveillance, and reconnaissance. Technical report, Space and Naval Warfare Systems Center, San Diego CA, 2001. 2

[40] G. Sun, T. Matsui, T. Kirimoto, Y. Yao, and S. Abe. Applications of infrared thermography for noncontact and noninvasive mass screening of febrile international travelers at airport quarantine stations. In *Application of Infrared to Biomedical Sciences*, pages 347–358. Springer, 2017. 2

[41] A. M. Waxman, M. Aguilar, D. A. Fay, D. B. Ireland, and J. P. Racamato. Solid-state color night vision: fusion of low-light visible and thermal infrared imagery. *Lincoln Laboratory Journal*, 11(1):41–60, 1998. 2

[42] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille. Adversarial examples for semantic segmentation and object detection. In *International Conference on Computer Vision. IEEE*, 2017. 5

[43] H. Yang, Q. Du, H. Su, and Y. Sheng. An efficient method for supervised hyperspectral band selection. *IEEE Geoscience and Remote Sensing Letters*, 8(1):138–142, 2011. 1

[44] Z. Ying, S. Simanovsky, R. Naidu, and S. Marcovici. Ct scanning systems and methods using multi-pixel x-ray sources, Aug. 22 2017. US Patent 9,739,724. 2

[45] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014. 4

[46] P. W. Yuen and M. Richardson. An introduction to hyperspectral imaging and its application for security, surveillance and target acquisition. *The Imaging Science Journal*, 58(5):241–253, 2010. 2