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]Department of Computer Science, The University of Texas at El Paso
[2]US Army Research Laboratory

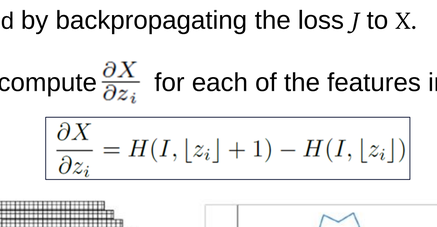
Introduction

- The remote sensing community has started to adopt deep learning and apply it to multispectral and hyperspectral imagery.
- 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.
- In addition, this high dimensionality can make the models highly susceptible to adversarially constructed inputs.
- 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.

Background

Previous approaches for high dimensionality reduction can be categorize into one of two types:

1. Transform data into a lower dimensional space:
   - Principal Component Analysis (PCA)
   - Linear Discriminant Analysis (LDA)



2. Band Selection:
   - Supervise Approaches
   - Unsupervised Approaches

ILFS

Learning in previous approaches follow a two steps procedure:

1. Select a subset of features.
2. Apply a learning algorithm using those selected features.

With ILFS dimensionality reduction is done simultaneously with learning a deep learning model to solve the learning task.

To discover and select the best features for the task, we compute:

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

Where,
I: Input space with n features.
Z: Vector encoding k selected feature $\{z_1, z_2, \ldots z_{k-1}, z_k\}$
X: Image that results from selecting features Z from I
J: Cost function of the network.

$\frac{\partial J}{\partial X}$ is computed by backpropagating the loss J to X.

For $\frac{\partial X}{\partial Z}$, we compute $\frac{\partial X}{\partial z_i}$ for each of the features in Z as:

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

Figure 1: H(*I;Z)* for a pixel by selecting *k* features

Experimental Setup

- DSTL Dataset:
  - 1 km x 1 km Satellite Image
  - Spatial resolution: 31 cm
  - 3 channels RGB
  - 8 Channels VNIR
  - 8 Channels SWIR
  - 10 Classes (Buildings, misc., roads, track, trees, crops, …)
  - DigitalGlobe's WorldView Satellite System
- Task: Semantic Segmentation
- Evaluation Metric: Mean IoU
- Architecture: Fully Convolutional Networks (FCN-8) with VGG-19 as backbone

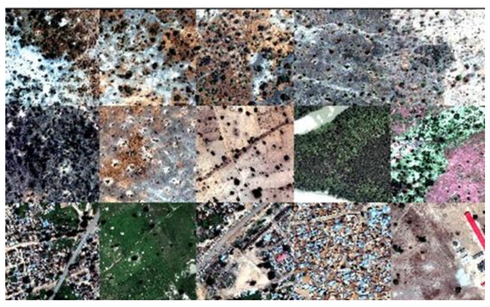


Figure 2: Some samples from DSTL Dataset

Performance Results

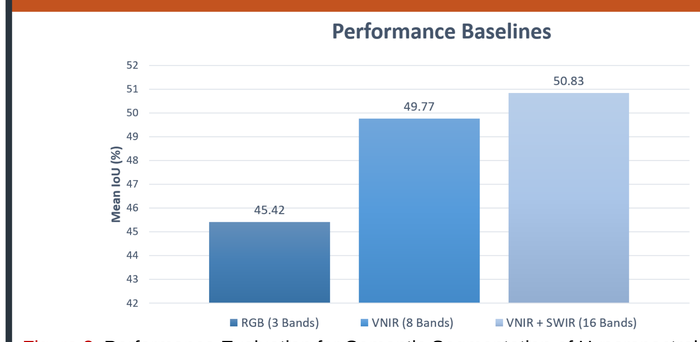


Figure 3: Performance Evaluation for Semantic Segmentation of Hyperspectral Images. All of models were trained on the same deep neural network architecture (VGG19-based FCN-8) for the same number of epochs. For RGB we trained using only the three RGB bands as input. For VNIR, we trained with all of the 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.

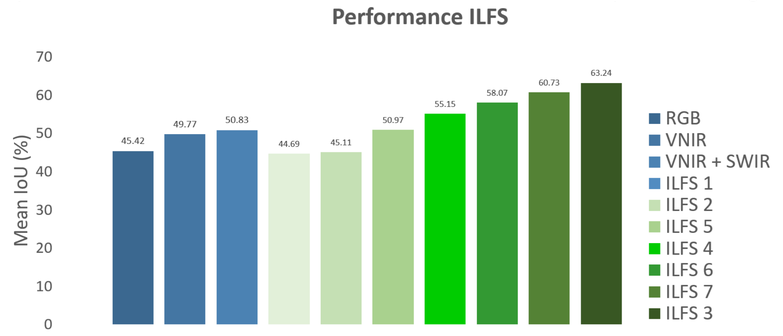
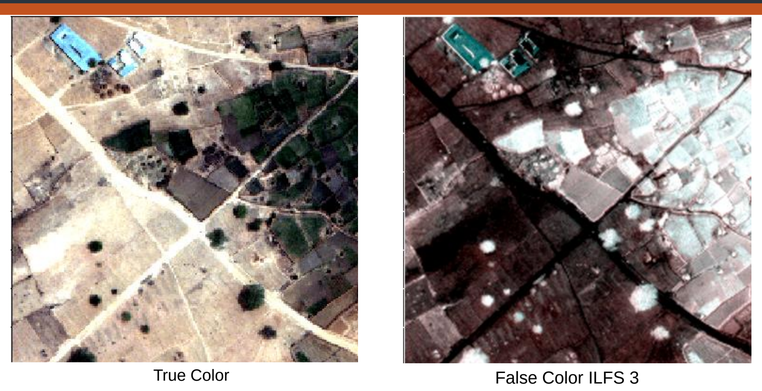


Figure 4: Performance evaluation for ILFS compared with Figure 3. ILFS 1...7 show the results when ILFS is used to pick 1, 2, 3, 4, 5, 6, and 7 input features, respectively. The results show that ILFS improves performance by up to 12.409% over the best model without ILFS. All ILFS models with more than 2 features beat all of the non-ILFS models from Figure 3.

ILFS Visualization



True Color | False Color ILFS 3

Figure 5: ILFS 3 Visualization. We see similar results to what is known as spectral indices in the remote sensing community, and can visually discriminate classes.

Adversarial Examples Beyond RGB

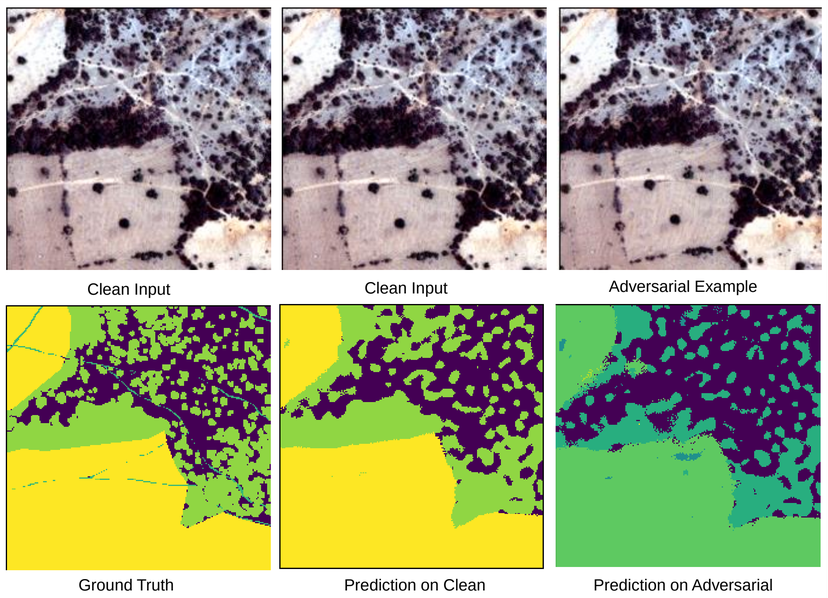


Figure 6: An adversarial example generated with l_∞-norm of 4 using FGSM.

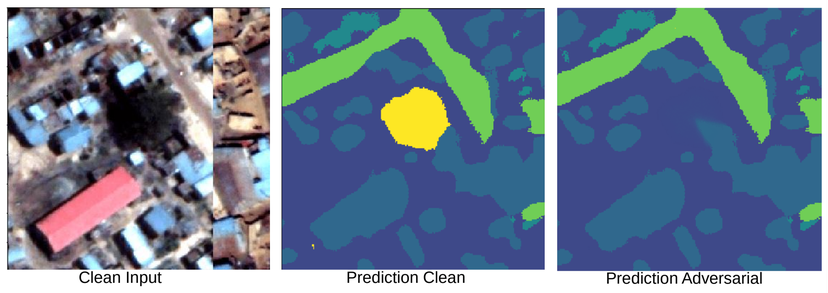


Figure 7: Dynamic Adversarial Perturbation for HSI Semantic Segmentation attack. This attack can be used to hide specific classes and/or objects from the scene while giving as output a prediction that is as close as possible to the real prediction. This figure shows how we successfully hid a tree for the adversarial prediction.

Spectral Signature Adversarial Examples

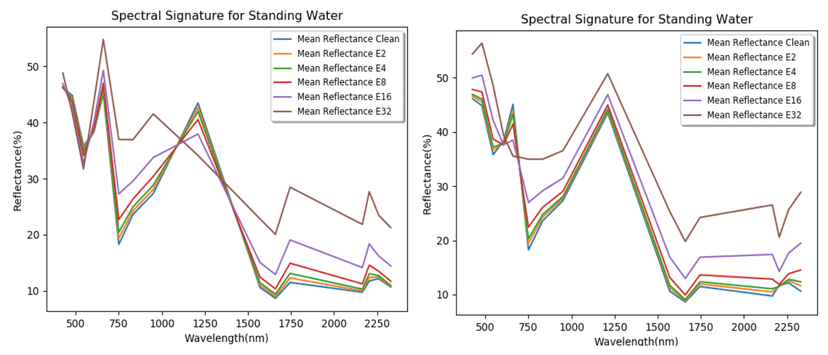


Figure 8: Spectral Signature of Adversarial Examples. Larger l_∞-norm produces more drastic changes to the spectral signature of the class.

Robustness Performance

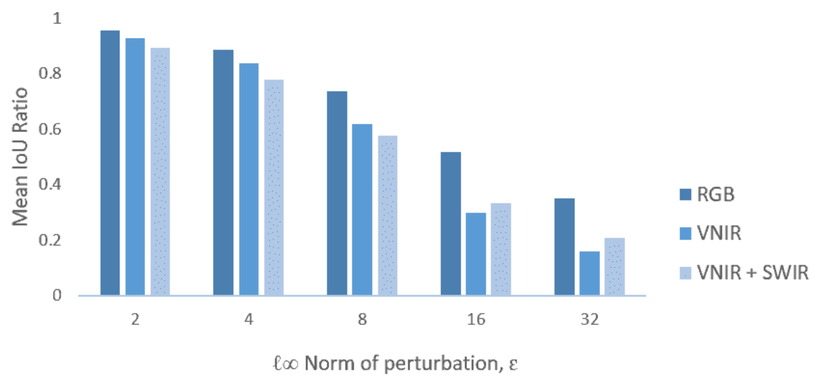


Figure 9: Robustness of multispectral image-based models to adversarial examples. We observe that models trained on high dimensional images are even more vulnerable to adversarial examples than RGB image-based models for white box settings for all four tested attacks.

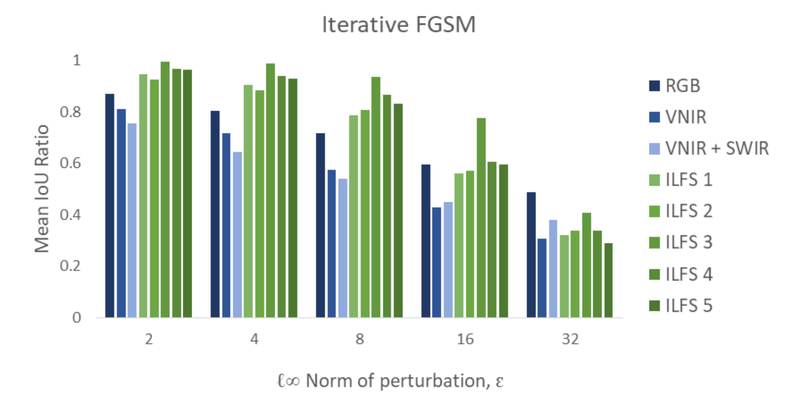


Figure 10: ILFS as a defense to adversarial examples. we attacked ILFS models obtained selecting different amount of features with four attack methods and different l_∞-norm of perturbation. ILFS models not only achieve better performance on clean inputs but are also more robust to adversarial examples as their mean IoU ratio is consistently higher.

ILFS Training Procedure Helps

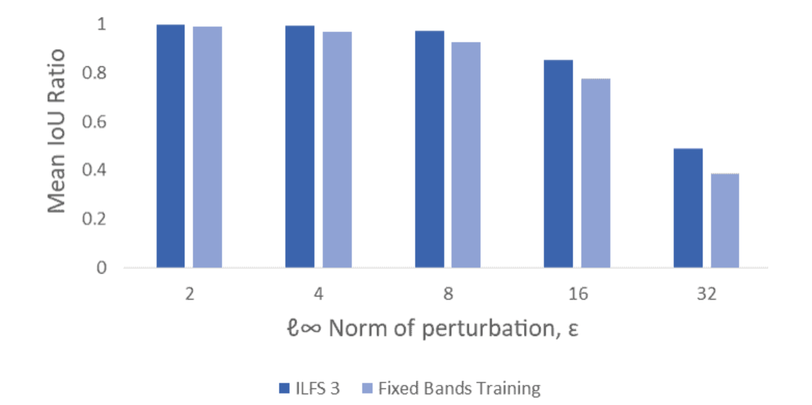


Figure 11: We trained a model using the output bands from ILFS 3 as a fixed input and compare it's robustness with the model trained using Integrated Learning and Feature Selection. Figure 11 shows that the model trained on fixed inputs is less robust. This supports our hypothesis that as ILFS keeps changing the input space during training, ILFS forces the network to perform well even when the input is maliciously modified.

Conclusions

- ILFS is effective for dimensionality reduction and significantly improves performance on the semantic segmentation task for high dimensional imagery.
- We showed that non-RGB machine learning models are vulnerable to adversarial examples.
- ILFS is an effective defense against adversarial examples on high dimensional imagery.

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. 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.
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. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. stat, 1050:19, 2017.
5. A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. International Conference on Learning Representations (ICLR), 2017.

Acknowledgements

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

14th IEEE Workshop on Perception Beyond the Visible Spectrum
PBVS 2018
In Conjunction with CVPR 2018 Salt Lake City
Salt Lake City, Utah, USA - June 18, 2018

DS3 DATA SCIENCE SUMMER SCHOOL