

Generating Natural Adversarial Hyperspectral examples with a modified Wasserstein GAN

Jean-Christophe Burnel^{1*}, Kilian Fatras¹, Nicolas Courty¹

University of South Brittany, CNRS, IRISA, UMR 6074, France

Abstract. Adversarial examples are a hot topic due to their abilities to fool a classifier’s prediction. There are two strategies to create such examples, one uses the attacked classifier’s gradients, while the other only requires access to the classifier’s prediction. This is particularly appealing when the classifier is not full known (black box model). In this paper, we present a new method which is able to generate natural adversarial examples from the true data following the second paradigm. Based on Generative Adversarial Networks (GANs) [5], it reweights the true data empirical distribution to encourage the classifier to generate adversarial examples. We provide a proof of concept of our method by generating adversarial hyperspectral signatures on a remote sensing dataset.

Keywords: Adversarial Examples · Remote sensing · Deep Learning.

1 Introduction

One of the weakness of DNNs is the so called adversarial example sensibility. Adversarial examples are samples which are not classified in the correct class. These examples can occur even for very accurate classifiers and can be problematic for their real-world applications, especially when security is an issue. Therefore one may want to have access to examples which can fool a classifier in order to measure how robust is the DNN.

Adversarial examples have received a lot of attention from the machine learning community. There are two main strategies to generate such examples. The first strategy is to use the gradients from classifiers. The generated adversarial examples are images drawn from the true distribution where we add a small perturbation following the gradients. The intuition is that the classifier will change its prediction because the gradients are the steepest classifier variation directions. This is the main idea of the Fast Gradient Sign Method (FGSM) algorithm [4]. One problem from this approach is that one needs to know the full architecture to compute the classifier’s gradients. The second paradigm does not require access to gradients. It looks for adversarial examples only by having access to the classifier’s prediction. Our new method follows the second strategy as we detailed later.

To generate adversarial examples without requiring to classifier’s gradients, [10] uses a GAN and an inverter. The generator is a function which goes from the latent space to the true data space while the inverter is an inverse function. From the true data, they use the inverter to go to the latent space and look for adversarial examples. They

* Contact author: jean-christophe.burnel@irisa.fr

add small perturbations to the latent vector z^* which represents a data x^* until they find an adversarial example. This method allows them to have very realistic adversarial examples without using the classifier’s gradients. In [9] authors use an AC-GAN to model the data distribution with the assumption that an ideal model could generate all the set of legitimate data. With such a model they search in the latent space for all the adversarial examples. Their search is done by minimizing the confidence score of a classifier while having the auxiliary classifier still predicting the correct class. So the adversarial examples look like true images and are adversarial for the attacked classifier.

We propose a different approach to generate natural adversarial examples from those methods. Like them, our method only needs to know the output of a given classifier, seen as it a ”black box” classifier. From a re-weighted distribution of the true data based on a classifier’s predictions, a generator is learnt to generate adversarial examples. The idea is to create a map between the latent space and the set of adversarial images which are present in the true data.

The paper is structured as follows: after a brief introduction on GAN, we present our **main contributions** in section 3. We will detail our modified loss function for the WGAN architecture which is specialised in adversarial examples generation. In section 4, we will test our method on hyperspectral data from the Data Fusion Contest 2018 [1], by generating data which are misclassified by a classifier.

2 Proposed methods

Generative Adversarial Networks. The principle of a GAN is to learn a generator of realistic data. Intuitively, it tries to minimize the distance between the distributions of generated data and target data. This notion of distance between distribution is critical and has led to several variants of GAN. Formally, the generator takes $z \sim \mathbb{P}_Z$, as input to generate data $G(z) \sim \mathbb{P}_G$ and then distribution their distance to real data distribution \mathbb{P}_r . \mathbb{P}_Z is usually a gaussian or a uniform distribution. GANs improve the generated data quality by minimizing the distance between the distributions \mathbb{P}_G and P_r . However as we do not know the true distributions \mathbb{P}_G and \mathbb{P}_r , we only consider their empirical distributions from the available true data and the generated ones. Wasserstein GAN [2] is a GAN variant which uses the Wasserstein distance to minimize the distance between distributions. Thanks to the Kantorovich-Rubinstein duality [8] for calculating this Wasserstein distance, the objective function of WGAN is :

$$\min_{\theta} \max_{\phi} \mathbb{E}_{x \sim \mathbb{P}_r} [\mathcal{D}_{\phi}(x)] - \mathbb{E}_{z \sim \mathbb{P}_z} [\mathcal{D}_{\phi}(G_{\theta}(z))]$$

where \mathcal{D} is within the set of 1-Lipschitz functions parameterized by θ . Analogous to GANs, we still call \mathcal{D} the ”discriminator” although it is actually a real-valued function. In order to respect the 1-Lipschitz constraint, [2] used a weight clipping trick for the discriminator’s weights during the optimization procedure. However this practice is far from being optimal and might lead to instability during optimization and poor minima. Another method [7] involves a gradient penalty and enforces the gradient to be less or equal to one. This is the variant of WGAN that we will consider in the rest of the paper:

$$\min_{\theta} \max_{\phi} \mathbb{E}_{x \sim \mathbb{P}_r} [\mathcal{D}_{\phi}(x)] - \mathbb{E}_{z \sim \mathbb{P}_z} [\mathcal{D}_{\phi}(G_{\theta}(z))] + \mathbb{E}_{\hat{x} \sim \mathbb{P}_{\hat{x}}} [(\max\{0, \|\nabla \mathcal{D}_{\phi}(\hat{x})\|_2 - 1\})^2]$$

where $\mathbb{P}_{\hat{x}}$ is the distribution of samples along the straight lines between a pair of points from \mathbb{P}_r and \mathbb{P}_G .

Generating adversarial data. For a fixed trained classifier on the true data, we aim to find a way to generate adversarial data with GANs. Instead of encouraging the generator to generate data close to the true data, we will encourage it to generate adversarial data which follows the true data distribution \mathbb{P}_r . Intuitively, the generator will be a map between the latent space and the adversarial data from the true distribution. We opted for a modification of the loss function of WGAN by considering a reweighting strategy of the empirical true distribution. Indeed, The generator generates data similar to the true data because of the true data empirical distribution which is $\frac{1}{N} \sum_{i=1}^n \delta_{x_i}$. Our main idea is to reweight a data weight according to the classifier’s prediction. If a data is misclassified, the resulting weight will be bigger than a correctly classified data. The resulting distribution must have the following form: $\sum_{i=1}^n p_i \delta_{x_i}$, where: $\sum_{i=1}^n p_i = 1$. Now we will describe and review the impact of different reweighting methods:

Uniform weight Here we are in the case of traditional WGAN, all data are given the same weight so there is no particular reason for the generator to produce adversarial examples, we can visualize it with figure 1a where we have a distribution where all points got the same weight $1/N$.

Hard weighting An intuitive way to generate adversarial example is to only consider misclassified data. In figure 1b, we see the data given the following weighting : $[1 - c(x)] * x$, where $c(x)$ is the classifier output. The downside of this weighting strategy is that we are left with few data for very accurate classifier. Unfortunately, GANs are data hungry and then this method is not efficient to generate adversarial examples.

Soft weighting Another weighting strategy is to use the prediction vector from the classifier. This can be done through a softmax function, defined as:

$$\text{softmax}(x, w) \triangleq \frac{\exp(w * [x_j - x_{max}])}{\sum_{i=0}^K \exp(w * [x_i - x_{max}])} \quad (1)$$

where x is our prediction vector and w is a temperature coefficient that controls the entropy of the resulting distribution. In the figure 1c we can see an example of weighting using the softmax function with a temperature $w = 8$.

To demonstrate the relevance of the different reweighting methods, we use a toy dataset to illustrate our approach from Figures 1a to 1f, where we represent data from a certain class with points, the classifier’s decision boundary with a line, so that points under the line are correctly classified and points over the line are misclassified.

As mentioned before, an important hyperparameter of our softmax function is the temperature coefficient. Its value allows controlling the weight for data with low prediction score. In the next figure, we use a factor of 5 for figure 1d and we can see that while misclassified data are given bigger weights the major part of the points still have a decent amount of weight. In figure 1e the weights for data too far from the classification line start fading due to their weights being low and finally in figure 1f we are vastly using the points along the classification line.

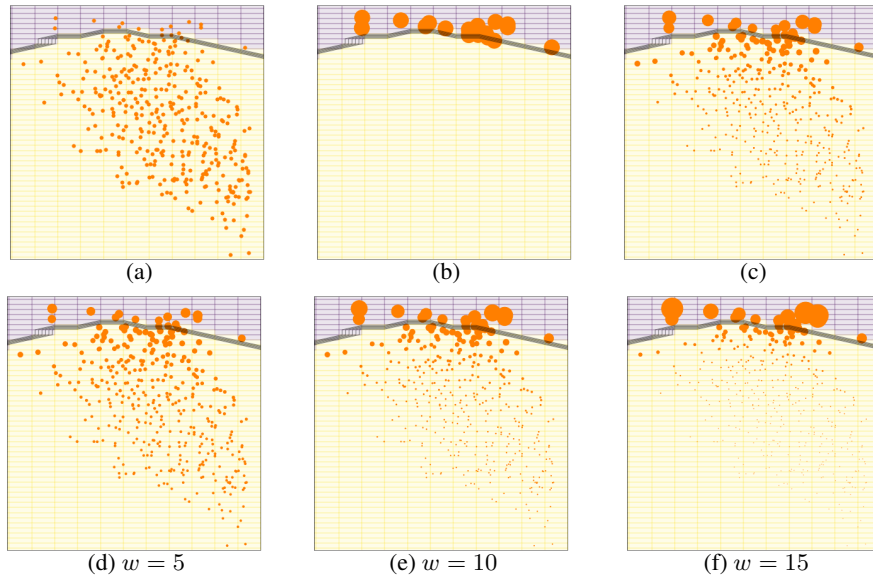


Fig. 1: Visualization of different weighting methods

3 Experiments and Results

Unlike traditional RGB images, hyperspectral imagery divides the color spectrum in multiple contiguous bands that can outreach the visible spectrum. In those images a pixel is a spectrum, and we can identify materials in a pixel by analysing it because each materials have its own spectral signature. To test our algorithm we first considered a pixel-based classification task on hyperspectral images. The classifier takes a spectrum as input and will output a probability vector of belonging to one particular material. The classifier we used is based on [6] and have an Overall Accuracy of 0.52 with $\kappa = 0.43$ on disjoint train/test, which seem coherent with the result observe in recent reviews [3]. Our objective here is to generate natural adversarial spectra.

Dataset. We use the DFC2018 dataset [1] which was acquired over the University of Houston campus and its neighbourhoods. It has hyperspectral data covering a 380-1050 nm spectral range with 48 bands at a 1-m GSD. Here the classes do not only cover urban classes (buildings, cars, railways, etc.) but it also has some vegetation classes (healthy or stressed grass, artificial turf, etc.). A view of the data is given with figure 2, where we only take 3 of the 48 bands for visualization purpose, and the ground truth labels are shown with figure 3

Experimentation. For this experiment we want to generate spectra of healthy grass, we know that the classifier has an accuracy of 82% for that particular class on the whole dataset. For the WGAN architecture we choose 1-dimensional convolutions for both the generator and the critic as it keeps coherence between close spectral bands. For the generator we have the followings layers (as seen in figure 4):

(Linear, 384)(Reshape, (6,64))(Conv1D, (12,32), 3)(Conv1D, (24,16), 3)(Conv1D, (48,1), 5), and for the critic we have :

Title Suppressed Due to Excessive Length



Fig. 2: False RGB

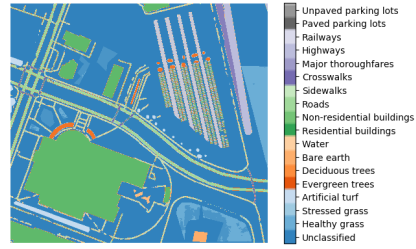


Fig. 3: Labels

(Conv1D, (24,16), 5)(Conv1D, (12,32), 3)(Conv1D,(6,64), 3)(Reshape, (384))(Linear, 1).
 We use all the available data. In this experiment we choose to use $w = 20$

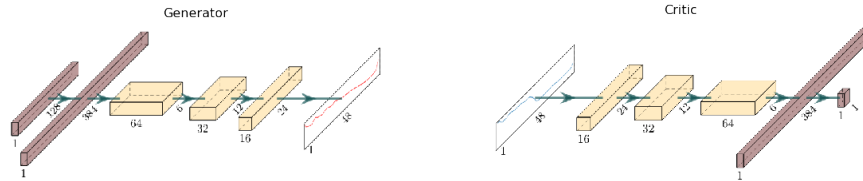


Fig. 4: Architecture used for Generator and Critic

Results. We consider the healthy grass for target class, and we want to generate spectra similar to it but misclassified by the classifier. For this we used our ground truth to replace the pixels of the target class by the closest generated spectra. To verify that our generated spectra belong to the target class, we choose to plot the statistics of the spectra, in plain the mean of the spectra and dotted the standard deviation. To interpret the results we choose to plot in the same figure the target class from the image, the adversarial examples generated and the two most predicted classes by our classifier. The later are evergreen trees and stressed grass figure 5. In both cases we see that the adversarial spectra match the targeted class more than the predicted class. With figure 6 we first show classification of healthy grass from the original image, here the classifier have 88.77% accuracy, and the classification of the adversarial spectra.

To go further we show some results with two other classes with different accuracy. We first choose Cars class where the attacked classifier have an accuracy of 0.39 and Crosswalks where the classifier have an accuracy of only 0.05. The results are visible with figure 7, here the comparison is made with the most predicted class, and we can see that it works in both cases. Moreover we show in table 1 that the more accurate our classifier is the less there are adversarial examples per batch. This can be explained by the fact that we do not have access to an infinite number of data so a more accurate classifier leads to less adversarial data to train the model.

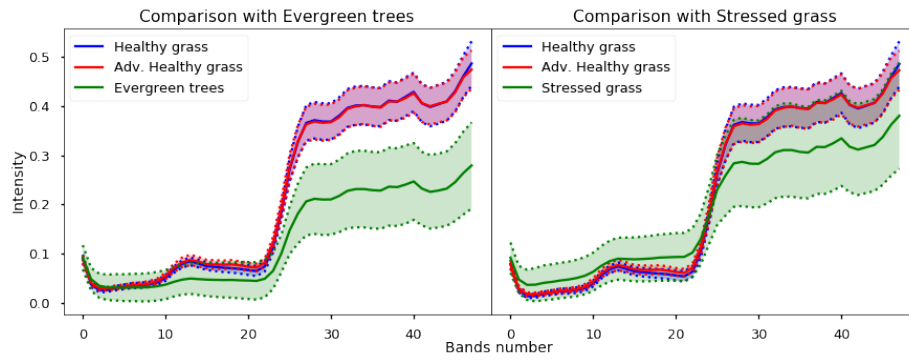


Fig. 5: [Best viewed in color] Comparison with Evergreen trees and Stressed grass

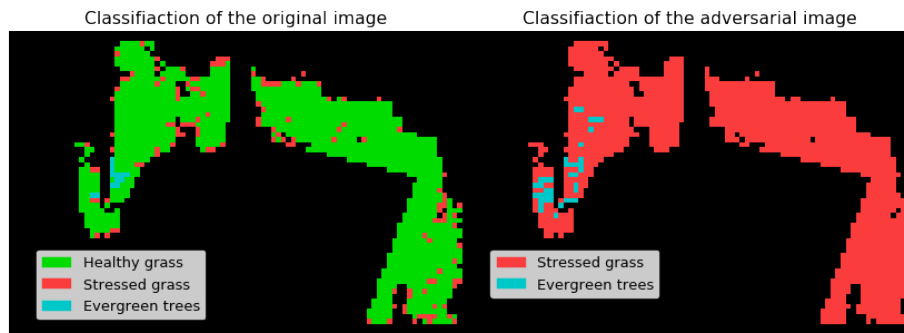


Fig. 6: [Best viewed in color] Comparison of predictions

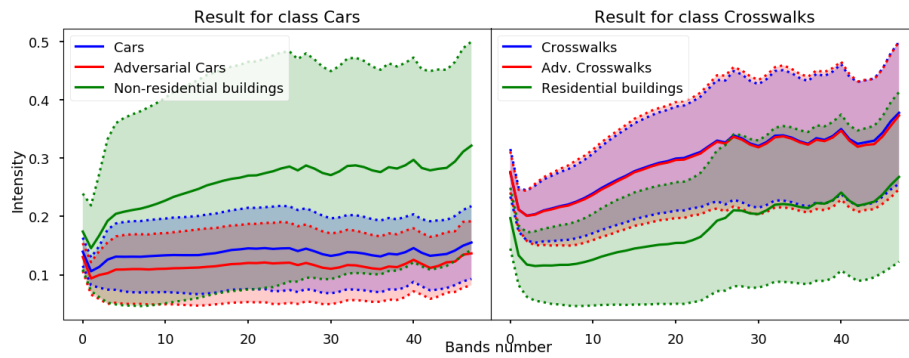


Fig. 7: [Best viewed in color] Results for classes Car and Crosswalks

Table 1: Success rate of our method (over 10 runs)

	Mean	Std	Clf. Acc
Healthy Grass	0.66	0.06	0.82
Car	0.81	0.05	0.39
Crosswalks	0.99	0.009	0.05

Qualitative comparison with state of the art. We now compare our results with [9]¹. We adapted this method to the same considered architectures considered in this paper. We used untargeted attacks, which means that we want our spectra to be misclassified but we do not want it to be classified as a specific label. However, it is to note that their algorithm is capable of doing targeted attack. The results are visible with figure 8. We did not include class Car for it was not possible to obtain any adversarial example over 100 runs of 10 000 iterations. For the Healthy grass class, the produced adversarial spectrum exhibits a lot of noise. For the class Crosswalks, it looks like their algorithm was not able to produce spectra close the real distribution, while our method worked in both cases. This somehow illustrates the benefits of our generation scheme. Further studies will try to confirm this view on a more extensive validation, involving practitioners from the field.

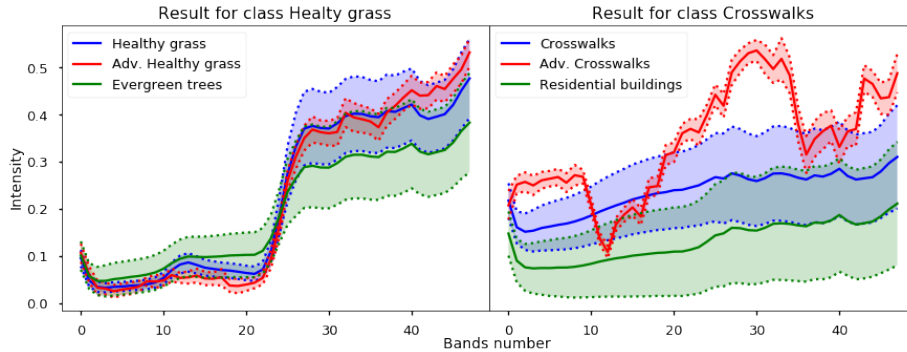


Fig. 8: [Best viewed in color] Results of [9] for classes Healthy grass and Crosswalks

4 Conclusion and discussion

In this work we propose a new method to generate natural adversarial examples for a pretrained classifier, using a modified loss function for the WGAN with a re-weighted distribution of our data based on the pretrained classifier’s prediction. Our method departs from other generative networks in the sense that we do not generate adversarial

¹ following their online implementation : https://github.com/ermongroup/generative_adversary

samples from existing ones by addition of a small perturbation. We rather consider the case where we train a generator to produce adversarial samples. As an immediate consequence, our method can treat the initial pretrained classifier as a black-box, which is a more realistic scenario.

We applied our method to the generation of hyperspectral data and show that it is possible to generate spectra that fool a classifier while keeping a similar content. One possible extension of our work would be to adapt our method to generate other modalities frequently encountered in remote sensing, such as point cloud data, but also to generate entire portions of images rather than solely focusing on spectra (spatial-spectral cubes). Here again, apart from the complexity of the needed neural architectures for such tasks, we do not see any major issue that would prevent our method from working. Complementarily, extensive tests with photo-interpreters will be conducted to further assess the qualities of our algorithm.

References

1. 2018 IEEE GRSS Data Fusion Contest. Online: <http://www.grss-ieee.org/community/technical-committees/data-fusion>
2. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: Proceedings of the 34th ICML. Proceedings of Machine Learning Research, PMLR (2017)
3. Audebert, N., Le Saux, B., Lefevre, S.: Deep learning for classification of hyperspectral data: A comparative review. *IEEE Geoscience and Remote Sensing Magazine* 7(2) (2019)
4. Goodfellow, I., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: International Conference on Learning Representations (2015)
5. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in neural information processing systems (2014)
6. Hu, W., Huang, Y., Wei, L., Zhang, F., Li, H.: Deep convolutional neural networks for hyperspectral image classification. *Journal of Sensors* **2015** (2015)
7. Petzka, H., Fischer, A., Lukovnikov, D.: On the regularization of wasserstein GANs. In: International Conference on Learning Representations (2018)
8. Peyré, G., Cuturi, M.: Computational optimal transport. *Foundations and Trends® in Machine Learning* **11**(5-6), 355–607 (2019)
9. Song, Y., Shu, R., Kushman, N., Ermon, S.: Constructing unrestricted adversarial examples with generative models. In: Advances in Neural Information Processing Systems (2018)
10. Zhao, Z., Dua, D., Singh, S.: Generating natural adversarial examples. In: International Conference on Learning Representations (2018)