SPARSE ATTACKS Papers Summary

Ritesh, Charan

January 2019

1 One Pixel Attack for Fooling Deep Neural Network

1.1 Introduction[1]

Main focus of adversarial examples is to fool the CNNs without getting into account of human perceptions. It can be done by adding quality perturbations to the pixels of images so that it can result into misclassification.But the problem arises when excessive perturbations is done to the image and it is not able to fool human perceptions. Hence, their comes a need to find the minimum amount of allowed perturbation which can be done to a image to fool the model with getting into eyes of humans. **Advantages[1] :**

- **Effectiveness**
- **Semi-Black Box Attack** : No particular attack function just focuses on the increasing of the probability of targeted class.
- **Flexibility** : No particular of differentiable or gradient designed neural networks

1.2 Methodology[1]

- **Problem Description :** f be a target image classifier receiving n-dimensional input, $\mathbf{x} = (x_1, x_2, ..., xn)$ be the original image in the form of vector correctly classified as t. $f_t(x)$ be the probability of $\mathbf{x}$ belonging to t. adv be the target class and e($\mathbf{x}$) = $(e_1, e_2, ..., e_n)$ is adversarial perturbation according to $\mathbf{x}$. Limit of maximum modification be L where L depends only on the length of vector e($\mathbf{x}$).The optimization function is as follows :

$$\begin{aligned} \max_{e(\mathbf{x})^*} \quad & f_{adv}(\mathbf{x} + e(\mathbf{x})) \\ \text{s.t.} \quad & ||e(\mathbf{x})|| \leq L \end{aligned} \tag{1}$$

  Problem arises in the above approach **determining which dimension we should perturb** and **how much should be the perturbation strength**. Here approach is slightly different :

$$\begin{aligned} \max_{e(\mathbf{x})^*} \quad & f_{adv}(\mathbf{x} + e(\mathbf{x})) \\ \text{s.t.} \quad & ||e(\mathbf{x})||_0 \leq d \end{aligned} \tag{2}$$

  where d is a small number denoting dimensions to be modified henceforth here d = 1. **Advantage of the above problem description is that we have no restriction on the making change in the strength of perturbation after fixing the dimensions as we don't have to focus on other dimensions perturbations.**

- **Differential Evolution :** It is a population based optimization method.In this algorithm during each iteration another set of solutions(children) is generated according to the current populations(parents). Then the children are compared with their corresponding parents,surviving if they are more fitted then their parents. In such a way DE is able to give us both efficiency and diversity in the solution.
  In DE we **require less information from target system** b/c in critical case their are

networks which are not differentiable and gradient based approach requires more information from the system.
Algorithm: Wasn't able to understand!

- **Method And Settings :** Explained in the code of one pixle attack on CIFAR-10 dataset.

1.3 Discussions[1]

- **Adversarial Perturbation:** Intially data points were moved in all directions with small changes which resulted into collectively large change on the natural image.But now we have specifie tthe direction already so we have to make considerable amount of change in that particular direction only. We can improve the accuracy of adverserial attacks by increasing the **iterations** and using improved DE or more advanced algorithm like **Co-variance Matrix Adaptation Evolution Strategy.**

- **Robustness of One Pixel Attack :** As on the use of low-cost algorithms,easy-implemented L_0 attack, we cannot expect one pixel attack to provide more robustness in comparison to other L_0 attacks.
  But in one sense one pixel attack is helpful as due to very less number of dimension changes it tooks detection methods a long amount of time to detect the adversarial perturbation which leads to increase in response time of the system and **sometime due to such low value of perturbation system doesn't take it into consideration.**

1.4 Future Works[1]

- Using of stronger algorithms like above mentioned ones
- Extension of DE to NLPs and Speech Recognition
- And many more technologies......

2 Sparse and Imperceivable Adversarial Attacks

2.1 Introduction[2]

This paper aims on modifying smallest amount of pixels in order to change the decision.This paper has the following key points :

- Suggesting a novel black-box attack based on local search which outperforms all existing l_0 attacks
- This paper present closed form expressions or algorithms for the projection onto l_0-ball in order to extend PGD attack
- Combining sparsity constraint and component-wise constraints and try to make attack more imperceivable.

This paper focuses on not changing colors in any particular direction so it becomes visible to humans by using locally constrained component-wise constraints.We find that adversarial training wrt l_2 partially decreases the effectiveness of l_0-attacks,while adversarial training wrt either l_2 or l_∞ helps to be more robust against sparse and imperceivable attacks.

2.2 Sparse and Imperceivable adversarial attacks

- **Sparse l_0-attack** : From practical point of view the l_0-attack tests basically how vulnerable the model is to failure of pixels or large localized changes on an object.
- **Sparse and Imperceivable Attack** : If we add more perturbation then a thresold value then it will be easily detected by human eyes. So in this attack we decide to put a bound on the amount of perturbations that can be added to the pixel.We have two specific goals :
  - We do not want to make changes along edges which are aligned with the coordinate axis as they can be easily spotted and detected.
  - We do not want to change the color too much and rather just adjust its intensity and keep approximately also its saturation level.

2.3 Algorithms for Sparse attacks

Algorithms for one-pixel modifications :

- **l_0-attack**
- **l_0+l_{inf}-attack**
- **$l_0 + sigmamap$-attack**

Multiple-pixels modifications :
Most of the times the modifications of one pixel are not sufficient to change the decision. **Let us assume we want to target an image to r class by changing at most k-pixels out of N-pixels. So we will proceed by choosing first k-perturbations among the N-one pixels perturbations according to a defined ordering π^r.** This ordering is done on the basis of an algorithm that encourages more effective change in pixels.

2.4 Projected Gradient Descent Algorithm

3 Adversarial Examples Are Not Bugs, They Are Features

3.1 Introduction

In this paper we see adversarial examples with new perspective.**This paper tries to prove that adversarial vulnerability is a direct result of our models sensitivity to well generalizing features in the data.** We find that our models learn to rely on these "non-robust" features, leading to adversarial perturbations that exploit this dependence.Explanation for **adversarial transferability** : Since any two models are likely to learn similar non-robust features, perturbations that manipulate such features will apply to both.For a model classification non-robust features are as much as important as robust features.
We try to show that it is possible to disentangle robust from non-robust features in standard image classification datasets.Specifically given any training datasets,we are able to construct :

- **A "robustified" version for robust classification**
- **A "non-robust" version for standard classification**

Both of the above versions are well-explained in the papers.

3.2 The Robust Features Model

We define **features** to be a funtion mapping from the input space X to the real numbers.
Key concepts regarding features :

- **ρ-useful features**: For a given distribution D, we call a feature f, ρ-useful($\rho \geq 0$)if it is correlated with the true label in expectation.
- **γ-robustly useful features**: Suppose we have ρ-useful feature f. We refer to f as a robust feature if, under adversarial perturbation(for some specified set of valid perturbations),f remains γ-useful.
- **Useful,non-robust features**: A useful, non-robust feature is a feature which is ρ-useful for some ρ bounded away from zero, but is not a γ-robust feature for any $\gamma \geq 0$. These features help with classification in the standard setting, but may hinder accuracy in the adversarial setting, as the correlation with the label can be flipped.

3.3 Finding Robust (and Non-Robust) Features

3.4 A Theoretical Framework for Studying (Non)-Robust Features

The conceptual framework of robust and non-robust features is strongly predictive of the empirical behaviour of state-of-the-art models on the real world. In order to further strengthen our understanding of the phenomenon, we instantiate the framework in a concrete setting that allows us to theoretically study various properties of the corresponding model.

- **The adversarial vulnerability can be explicitly expressed as a difference between the inherent data metric and the l_2 metric.**
- **Robust learning corresponds exactly to learning a combination of these two metrics.**
- **The gradients of adversarially trained models align better with the adversary's metric.**

3.5 Conclusion

In this work, we cast the phenomenon of adversarial examples as a natural consequence of the presence of highly predictive but non-robust features in standard ML datasets.We support it by explicitly disentangling robust and non-robust features in standard datasets, as well as showing that non- robust features alone are sufficient for good generalization. **Adverserial Examples are fundamentally human phenomenon**.
Classifiers exploit highly predictive features that happen to be non-robust under a human-selected notion.In similar manner as long as models rely on the non-robust features we cannot expect to have model explanations that are both human-meaningful and faithful to the model themselves.

4 Robustness May Be at Odds with Accuracy

4.1 Introduction

One can often synthesize small, imperceptible perturbations of the input data and cause the model to make highly-confident but erroneous predictions.This problem of so-called adversarial examples has garnered significant attention recently and resulted in a number of approaches both to finding these perturbations, and to training models that are robust to them. But creating these adversarially robust models have been proved quite challenging.

By the vulnerability of models trained using standard methods to adversarial perturbations makes it clear that the adverserially trained models are different to that of standard trained model.**Robustness comes at a cost**. Cost are in the form of expensive computaionally training methods. **The question arised in this paper is :** ***Are these only costs of adversarial robustness?*** **and if so, if we choose to pay these costs,*****would it always be preferable to have a robust model instead of a standard one?***

4.2 On the Price of Adversarial Robustness

Adversarial Robustness : In particular, there has been a lot of interest in developing models that are resistant to them, or, in other words, models that are adversarially robust. In this context, the goal is to train models with low expected adversarial loss.

Adversarial Training : Most Successfull approach for generation of adversrially robust models. Adversarial training is motivated by solving the corresponding(adversarial) empirical risk minimization.Though adversarial training is effective, this success comes with certain drawbacks. The most obvious one is an increase in the training time (we need to compute new perturbations each parameter update step).Another one is the potential need for more training data.All of this leads to increase in training of robust models and raise the question : **Are robust classifiers better than standard ones in every other aspect?**

Adversarial Training as a Form of Data Augmentation : A key implication of this view is that adversarial training should be beneficial for the standard accuracy of a model.When classifiers are trained with relatively few samples.In this setting, the amount of training data available is potentially insufficient to learn a good standard classifier and the set of adversarial perturbations used "compatible" with the learning task.In such regime, robust training does indeed act as data augmentation, regularizing the model and leading to a better solution.Surprisingly however, as we include more samples in the training set, this positive effect becomes less significant. In fact, after some point adversarial training actually decreases the standard accuracy. Now, the main question arises : **Why does there seem to be a trade-off between standard and adversarially robust accuracy?**

4.2.1 Adversarial robustness might be incompatible with standard accuracy

4.2.2 The Importance of Adversarial Training

In the distributional model D,a classifier that achieves very high standard accuracy will have near-zero adversarial accuracy.Hence, in an adversarial setting, where the goal is to achieve high adversarial accuracy, the training procedure needs to be modified.

Adversarial Training matters Theorem : For $\eta \geq 4/\sqrt{d}$ and p ≤ 0.975 (the first feature is not perfect), a soft-margin SVM classifier of unit weight norm minimizing the distributional loss achieves a standard accuracy of $\geq 99\%$ and adversarial accuracy of $\leq 1\%$ against an l_∞ -bounded adversary of $\epsilon \geq 2\eta$. Minimizing the distributional adversarial loss instead leads to a robust classifier that has standard and adversarial accuracy of p against any $\epsilon \leq 1$.

This theorem shows that if our focus is on robust models, adversarial training is necessary to achieve non-trivial adversarial accuracy in this setting. Soft-margin SVM classifiers and the constant 0.975 are chosen for mathematical convenience. Our proofs do not depend on them in a crucial way and can be adapted, in a straightforward manner, to other natural settings, e.g. logistic regression.

Transferability : An interesting implication of our analysis is that standard training produces classifiers that rely on features that are weakly correlated with the correct label. This will be true for any classifier trained on the same distribution. Hence, the adversarial examples that are created by perturbing each feature in the direction of y will transfer across classifiers trained on independent samples from the distribution. **Empirical examination :** Interestingly, we observe a qualitatively similar behavior.We see that the standard classifier assigns weight to even weakly-correlated features. The robust classifier on the other hand does not assign any weight beyond a certain threshold. Further, we find that it is possible to obtain a robust classifier by directly training a standard model using only features that are relatively well-correlated with the label (without adversarial training).

4.3 Unexpected Benefits of Adversarial Robustness

At a high level, robustness to adversarial perturbations can be viewed as an invariance property that a model satisfies.s. Thus, robust training can be viewed as a method to embed certain invariances in a model. Since we also expect humans to be invariant to these perturbations (by design, e.g. small l_p-bounded changes of the pixels), robust models will be more aligned with human vision than standard models.

- **Loss gradients in the input space align well with human perception**
- **Adversarial examples exhibit salient data characteristics**
- **Smooth cross-class interpolations via gradient descent**

Above points are well explained and highlighted in the research paper.

4.4 Conclusions

Specifically, we identify a trade-off between the standard accuracy and adversarial robustness of a model, that provably manifests even in simple settings. This trade-off stems from intrinsic differences between the feature representations learned by standard and robust models. Our analysis also potentially explains the drop in standard accuracy observed when employing adversarial training in practice.

Robust models learn meaningful feature representations that align well with salient data characteristics. The root of this phenomenon is that the set of adversarial perturbations encodes some prior for human perception. Thus, classifiers that are robust to these perturbations are also necessarily invariant to input modifications that we expect humans to be invariant to. We demonstrate a striking consequence of this phenomenon: robust models yield clean feature interpolations similar to those obtained from generative models such as GANs.This emphasizes the possibility of a stronger connection between GANs and adversarial robustness.Finally findings show that the interplay between adversarial robustness and standard classification might be more nuanced that one might expect.

5 Exploring the Landscape of Spatial Robustness

5.1 Introduction

In this we had shown that neural network–based vision classifiers are vulnerable to input images that have been spatially transformed through small rotations, translations, shearing, scaling, and other natural transformations. Such transformations spread widely in vision applications and hence quite likely to naturally occur in practice. the vulnerability of neural networks to such transformations raises a natural question : **How can we build spatial robust classifiers?** While these transformations appear natural to a human, we show that small rotations and translations alone can significantly degrade accuracy. For fine-grained understanding of the spatial robustness of standard image classifiers for the different datasets following properties were taken into account :

- **Classifier Brittleness**
- **Relative adversary strength**
- **Spatial loss landscape**
- **Improving spatial robustness**
- **Combining spatial and l-bounded attacks**

5.2 Adversarial Rotations and Translations

An adversarial example for a given input image x and a classifier C is an image x* that satisfies two properties :

- label generated for x* is different then x i.e. C(x*)=C(x)
- the adversarial example x* is **"visually similar"** to x

The two images are indeed visually similar when they are close enough in some l_p-norm. However, the converse is not necessarily true. A small rotation or translation of an image usually appears visually similar to a human, yet can lead to a large change when measured in an l_p-norm. We aim to expand the range of similarity measures considered in the adversarial examples literature by investigating robustness to small rotations and translations.
Attack methods : Our first goal is to develop sufficiently strong methods for generating adversarial rotations and translations. In the context of pixel-wise l_p-bounded perturbations, the most successful approach for constructing adversarial examples so far has been to employ optimization methods on a suitable loss function.We implement this transformation in a differentiable manner using the spatial transformer blocks. In order to handle pixels that are mapped to non-integer coordinates, the transformer units include a differentiable bilinear interpolation routine. Since our loss function is differentiable with respect to the input and the transformation is in turn differentiable with respect to its parameters, we can obtain gradients of the model's loss function w.r.t. the perturbation parameters. This enables us to apply a first-order optimization method to our problem.
We compute the perturbation in three distinct ways:

- **First-Order Method(FO) :** Starting from a random choice of parameters, we iteratively take steps in the direction of the gradient of the loss function. This is the direction that locally maximizes the loss of the classifier. Since the maximization problem we are optimizing is non-concave, there are no guarantees for global optimality, but the hope is that the local maximum solution closely approximates the global optimum.
- **Grid-Search :** We discretize the parameter space and exhaustively examine every possible parametrization of the attack to find one that causes the classifier to give a wrong prediction (if such a parametrization exists). Since our parameter space is low-dimensional enough, this method is computationally feasible (in contrast to a grid search for l_p-based adversaries).

- **Worst-of-k :** We randomly sample k different choices of attack parameters and choose the one on which the model performs worst. As we increase k, this attack interpolates between a random choice and grid search.

A first-order attack requires full knowledge of the model to compute the gradient of the loss with respect to the input, the other two attacks do not. They only require the outputs corresponding to chosen inputs, which can be done with only query access to the target model.

5.3 Improving Invariance to Spatial Transformations

Augmenting the training set with random rotations and translations does improve the robustness of the model against such random transformations. However, data augmentation does not significantly improve the robustness against worst-case attacks and sometimes leads to a drop in accuracy on unperturbed images. To address these issues, we explore two simple baselines that turn out to be surprisingly effective.

Robust Optimization : Instead of performing standard empirical risk minimization to train the classification model, we utilize ideas from robust optimization.The main barrier to applying robust optimization for spatial transformations is the lack of an efficient procedure to compute the worst-case perturbation of a given example. Performing a grid search is prohibitive as this would increase the training time by a factor close to the grid size, which can easily be a factor 100 or 1,000.
Given that we cannot fully optimize over the space of translations and rotations, we instead use a coarse approximation provided by the worst-of-10 adversary. So each time we use an example during training, we first sample 10 transformations of the example uniformly at random from the space of allowed transformations. We then evaluate the model on each of these transformations and train on the one perturbation with the highest loss. This corresponds to approximately minimizing a min-max formulation of robust accuracy.

Aggregating Random Transformations : The accuracy against a random transformation is significantly higher than the accuracy against the worst transformation in the allowed attack space. This motivates the following inference procedure: compute a (typically small) number of random transformations of the input image and output the label that occurs the most in the resulting set of predictions. We constrain these random transformations to be within 5% of the input image size in each translation direction and up to 15 degree of rotation. The training procedure and model can remain unchanged while the inference time is increased by a small factor.

Combining both the Methods : Above two mentioned methods are orthogonal.We can therefore combine robust training (using a worst-of-k adversary) and majority inference to further increase the robustness of our models.

5.4 Conclusions

We examined the robustness of state-of-the-art image classifiers to translations and rotations. We observed that even a small number of randomly chosen perturbations of the input are sufficient to considerably degrade the classifier's performance.
Also, our results underline the need to consider broader notions of similarity than only pixel-wise distances when studying adversarial misclassification attacks. **In particular, we view combining the pixel-wise distances with rotations and translations as a next step towards the "right" notion of similarity in the context of images.**

6 References

1. Jiawei Su*, Danilo Vasconcellos Vargas* and Kouichi Sakurai. **One-Pixel-Attack-for-Fooling-Deep-Neural-Networks**.

2. Francesco Croce, Matthias Hein. **Sparse and Imperceivable Adversarial Attacks**

3. Andrew Ilyas*,Shibani Santurkar*,Dimitris Tsipras*,Logan Engstrom*,Brandon Tran,Aleksander Madry **Adversarial Examples Are Not Bugs, They Are Features**

4. Shibani Santurkar*,Dimitris Tsipras*,Logan Engstrom*,Alexander Turner,Aleksander Madry **Robustness May Be at Odds with Accuracy**

5. Logan Engstrom*,Brandon Tran*,Dimitris Tsipras*,Ludwig Schmidt,Aleksander Madry **Exploring the Landscape of Spatial Robustness**