Effectiveness of adversarial examples on class-imbalanced Convolutional Neural Networks

Rafael Possas, Ying Zhou

University Of Sydney, Camperdown, NSW
{rafael.possas,ying.zhou}@sydney.edu.au
https://sydney.edu.au

Abstract. A considerable amount of literature was published on the performance of Convolutional Neural Networks and the field has evolved considerably in the last years. However, as most machine learning methods, these networks suffer from the data imbalance problem - when the underlying training dataset is comprised of unequal number of samples for each label/class. Such difference naturally causes a phenomenon known as domain shift, which can be explained by the low generalisation capabilities of a model when presented with previously unseen data. Recent research have focused on a technique called Gradient Sign that forces domain shift on deep networks by creating adversarial examples. These are usually comprised of small directed changes on original data points that causes inputs to be misclassified by the predictive algorithm. Recent developments in such methods have heightened the need for better understanding of this phenomena. This study focuses on an experimental approach that sheds light on the link between the imbalanced learning problem and adversarial examples. Through a series of experiments we evaluated the gradient sign methods on imbalanced datasets linking their effectiveness with the underlying data distribution of image recognition models.

Keywords: convolutional neural networks, adversarial examples, gradient sign, imbalanced training

1 Introduction

To date, little evidence has been found associating training data characteristics of a model with its robustness to adversarial examples. Experimental demonstrations of adversarial effects were carried out mainly by [3], [6], [11], [14]. The motivation for Adversarial robustness comes largely from being able to shield image recognition systems from behaving unexpectedly. Previous published studies are limited to showing the general effectiveness of adversarial methods rather then understanding the deep relationship with the underlying training data distribution.

This work presents a practical approach for evaluation of gradient sign methods on datasets with skewed distributions. The main contributions of this work are as follows:

1. Understanding of the relationship between the gradients learned on an image recognition model and its vulnerability to gradient sign methods
2. Evaluation of the resilience of imbalanced CNNs to adversarial attacks using same model knowledge and transfer learning from different models.
3. Investigation of classes with similar distributions and their relationship to both adversarial attacks and imbalanced learning problem.

Section 2 of this paper discusses the related work in both convolutional neural networks and gradient sign methods. Section 3 provides details of the training models, imbalanced datasets and gradient sign methods used in our experiments. Section 4 presents the results on the under-sampled, over-sampled and balanced cases using both same/different model gradient. Sections 5 and 6 are dedicated to drawing conclusions and providing directions to related future work.

2 Related Work

Recent work has shown that the generalisation capabilities of deep networks is rather sparse [11], [13]. Thus, there is an opportunity for methods to exploit empty pockets of space and, hence, systematically create an adversary that produces an incorrect, high confidence prediction for an image without visible distortion. This can be done by adding just enough intentional noise to each pixel of the image so as to fool an algorithm into thinking that the image has an incorrect label [6], [11], [13], [16].

The Gradient Sign method developed by Goodfellow et al. (2014) has been used as the foundation of many of the experiments in adversarial crafting on CNNs. The results have shown that the model can possibly have linear behavior in very high dimensional spaces. Most inputs were miss-classified not only by Goodfellow et. al [6] experiments but many others [3],[11],[14].

Developing robustness to adversarial examples has already been approached as academic work [3], recent research has shown that one can use transfer learning to perform black-box attacks against Deep Neural Networks [17] , [14] .The combination of adversaries and transfer learning creates a threat vector for many state of the art methods. Attacks can, however, depend on some specific internal information of the target model [13]. As most recent applied methods depend on the network gradient information, there is a straight dependence on the network confidence per class label and the robustness of the model to adversarial attacks.

Imbalanced learning is a well known cause for lower performance of several machine learning algorithms [7], [9]. Data distribution on real world is often skewed and rarely contains enough information to learn all the required features of the data domain. Adversaries are proven to explore class distributions vicinities, and the question whether imbalanced training sets affects their effectiveness is still unanswered.

3 Initial Setup

For the purposes of this work, we used a dataset that is not only visually rich but also enables analysis between different class labels. The 2014 ImageNet dataset [5] would be the natural choice, however, the amount of classes (1,000) would make it harder to perform comparisons. We use a dataset with similar characteristics, the CIFAR-10 [10]. It contains 10 different class labels of 32x32 images, which enables us to easily train our algorithm from scratch and also to understand deeper relationships between labels.

3.1 Network Architecture and parameters

All the experiments were done using the VGG architecture [15]. The two FC-4096 layers at the end were replaced by one single layer with 512 neurons and RELU activations. In addition, the total number of convolutions blocks and pooling were reduced to 3, with the first layer having 2 stacked convolution layers followed by a max pooling of stride 2x2 and the last two layers with 3 stacked convolutions also followed by a max pooling of stride 2x2. We have used RMSProp [2] as the optimisation technique with a learning rate of 10^{-4} and the decay 10^{-5}. Figure 1 shows the results of our models on the CIFAR10 original dataset.

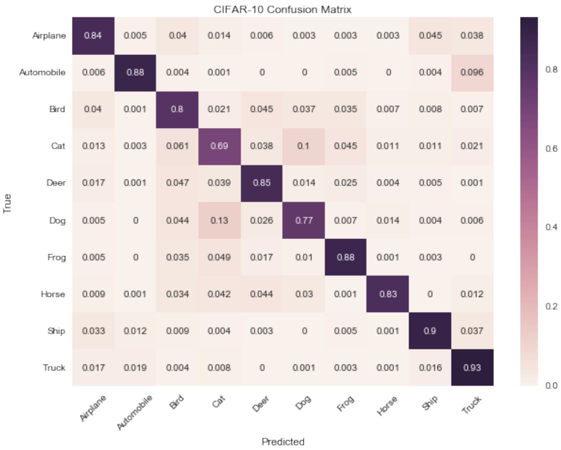
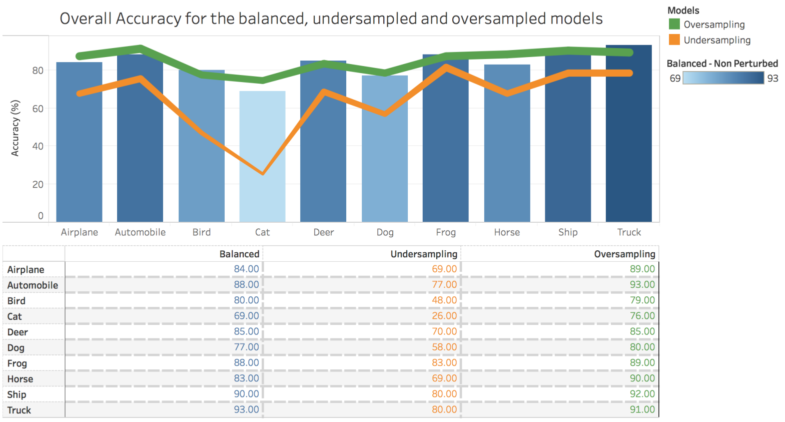


Fig. 1. Results of our adapted VGG architecture on CIFAR-10 dataset

3.2 Synthetic data imbalance

As the CIFAR10 dataset is not naturally imbalanced, we have artificially created two variations on which we trained our networks. While one dataset consisted on a direct under-sample of the target class to 1,000 samples, the other was changed using an oversampling of the target class(or an under-sampling of all other classes). We kept the number of samples for the target class at 5,000 while all other classes were reduced to 1,000 samples. For each class of the two different datasets configurations, a network was then trained until convergence using the same hyper-parameters as the balanced case. Each model was evaluated against a test set of 10,000 equally distributed samples with the target class being perturbed by its own under/over-sampled model and the balanced model. Both imbalanced networks were separately tested for each class for both same model and different model adversaries. The same model test aims to understand the vulnerability of class imbalance on adversarial examples while the different model test main goal is to verify the robustness on transfer learning environments. In total we evaluated 50 different combinations: 20 for each different imbalanced dataset (same model gradient and balanced network gradient) and 10 for the balanced network using its own gradients on each class. Figure 2 shows the accuracy for the models without any perturbation.



	Balanced	Undersampling	Oversampling
Airplane	84.00	69.00	89.00
Automobile	88.00	77.00	93.00
Bird	80.00	48.00	79.00
Cat	69.00	26.00	76.00
Deer	85.00	70.00	85.00
Dog	77.00	58.00	80.00
Frog	88.00	83.00	89.00
Horse	83.00	69.00	90.00
Ship	90.00	80.00	92.00
Truck	93.00	80.00	91.00

Fig. 2. Target class accuracy on all models

3.3 Gradient Sign methods

The Gradient Sign is a method that uses internal gradient information so as to create directed perturbation to test data. Whether one adds or subtract noise according to equations 1 and 2, the resulting label will have different meanings. Suppose the current true label of the class is selected as a gradient candidate, adding noise would mean that we increase the cost function of our input while subtracting noise is the same as minimizing our loss even further. These approaches will from now on be referred as ascent and descent methods. Perturbations could also be applied by two different variations of the gradient sign

[6]. While the Fast Gradient Sign applies a single perturbation to the input, the Iterative Gradient Sign performs the same perturbation a chosen number of times iteratively. Figure 3 shows an example of adversarial created using the fast method.

$$C(x + \delta) \approx C(x) + \epsilon * sign(\nabla C) \tag{1}$$

$$C(x + \delta) \approx C(x) - \epsilon * sign(\nabla C) \tag{2}$$

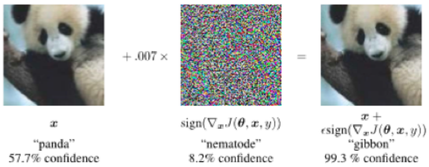

+ .007 ×
=
$\boldsymbol{x}$
"panda"
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

Fig. 3. Adversarial example crafting with fast gradient sign [6].

3.4 Backpropagated Class Gradient

The choice of the backpropagated class gradient has direct influence on the generated adversary. In order to reduce the variability of our experiment, we have chosen the true sample label as the backpropagated gradient coupled with the ascent method. The intuition behind this choice is that we look to increase the cost function of the target class by moving away from the current true label. Moreover, in order to test our networks, we needed an ϵ value that would not only keep the image meaning understandable to human perception but also provide only the minimum amount of perturbation so as to push most of the samples to the closest vicinity leading to a successful misclassification. From all the trials performed, the value of ϵ that seemed to fulfil our needs was 0.01.

4 Results

The baseline of our comparison is done through the performance of a balanced network to adversarial attacks. Table 1 shows that the accuracy for all classes is drastically reduced when the balanced model is presented with adversarial examples. The effectiveness of the adversarial attack can be partially explained by the balancing of the dataset . In a model where the dataset used in training aims for normalization over all classes, the network is often caught in trying to find weights and biases that generalizes well over all set of labels. Therefore, small perturbations become more efficient due to a bigger proximity of classes distributions in space.

4.1 Class under-sampling and over-sampling

Networks with under-sampled datasets were more vulnerable when presented with adversarial examples. Figure 5 shows the relative difference for all the three networks (balanced, under-sampled and over-sampled). Values were calculated by finding the difference between the perturbed accuracy and the non-perturbed accuracy of each class model. They represent the percentage on which the initial accuracy was reduced. The under-sampled model had the higher relative difference on average, which shows that the imbalanced nature of the dataset ended-up increasing the vulnerability of the model.

Class imbalanced models are naturally affected by the false positive and false negative trade off shown on figure 4. The decision boundaries on such models favour the class with more samples and, hence, increases the accuracy for one class while decreasing for the other classes. The area under the curve for misclassified examples on the under-sampled distribution is bigger, and it is caused by the suboptimal exploration of feature space of that class. This effect is exploited by adversaries as there is an increase on the misclassification rate of distributions with lower amplitude. An under-sample of a specific label causes its distribution to be squished into space and, hence, have less impact on the definition of decision boundaries.

	Different Model		Same Model		
Class Label	Undersample	Oversample	Balanced	Undersample	Oversample
0 - Airplane	60%	87%	36%	19%	61%
1 - Automobile	64%	91%	23%	16%	63%
2 - Bird	38%	73%	20%	9.4%	27%
3 - Cat	21%	72%	11%	0.5%	19%
4 - Deer	58%	80%	20%	9.8%	20%
5 - Dog	47%	76%	15%	9%	38%
6 - Frog	76%	88%	27%	20%	49%
7 - Horse	59%	88%	20%	18%	52%
8 - Ship	69%	89%	37%	19%	59%
9 - Truck	46%	87%	49%	21%	54%

Table 1. Results for the two different sources of perturbations along with the two different imbalanced datasets

Perturbation on the over-sampling case had a weaker effect, as the small push caused by our ϵ was not enough to move points to outside of their distributions. Objects of the over-sampled classes would need bigger steps in order to successfully create an adversarial that leads to a wrong classification label. Accuracy for most of the over-sampling cases was around 45% and the relative difference was the lowest of all three models, which shows robustness of the target over-sampled class.

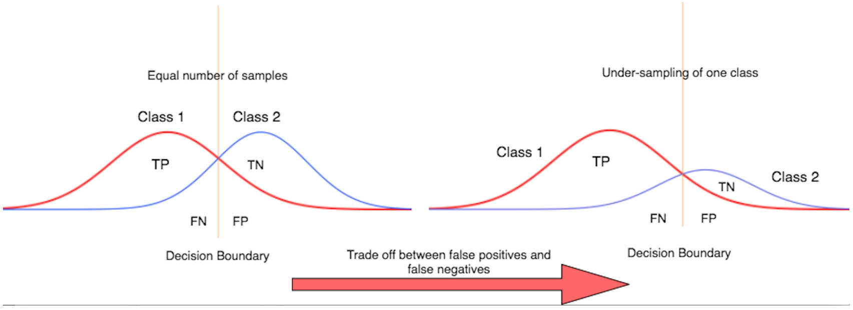
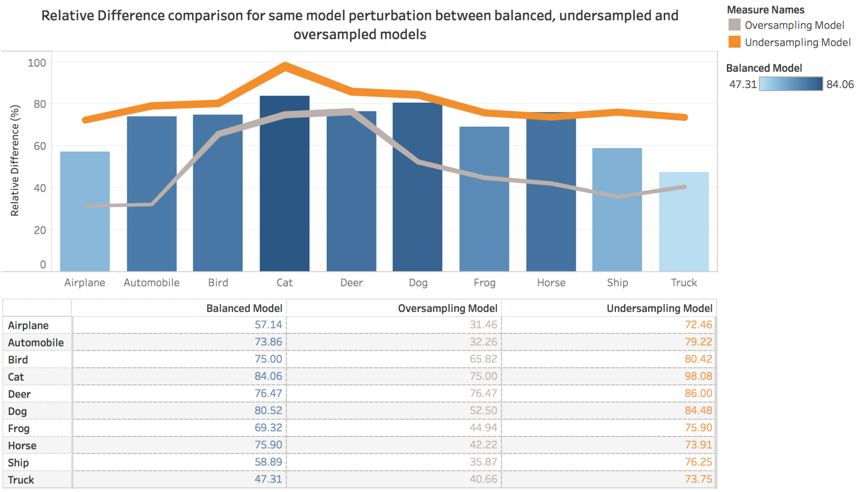


Fig. 4. Dataset imbalance causes models to perform adjustments of decision boundaries leading to an increase on accuracy of the majority class and decrease on the minority class.



	Balanced Model	Oversampling Model	Undersampling Model
Airplane	57.14	31.46	72.46
Automobile	73.86	32.26	79.22
Bird	75.00	65.82	80.42
Cat	84.06	75.00	98.08
Deer	76.47	76.47	86.00
Dog	80.52	52.50	84.48
Frog	69.32	44.94	75.90
Horse	75.90	42.22	73.91
Ship	58.89	35.87	76.25
Truck	47.31	40.66	73.75

Fig. 5. Relative difference for each model. Higher numbers means more vulnerability

The increased number of samples of the over-sampled label causes the network to perform a trade-off when optimizing its loss function. For instance, the decision boundary would be chosen in order to minimize the total error of the network. The cost function is lower when the decision boundary minimizes the misclassification of the majority class as there is a higher number of samples. The choice of a biased decision boundary could be one of the factors explaining the higher resilience of over-sampled networks.

4.2 Transfer Learning

The use of a different model for creating adversaries has shown less effective when compared to the same model attack. As the overall gradient have not only different direction but also magnitudes, the attacked system has proven to be more robust. The experiment reveals that although Gradient Sign is quite effective for fooling networks it does require a good amount of knowledge from the underlying training parameters so as to unleash its full potential. Attacking an

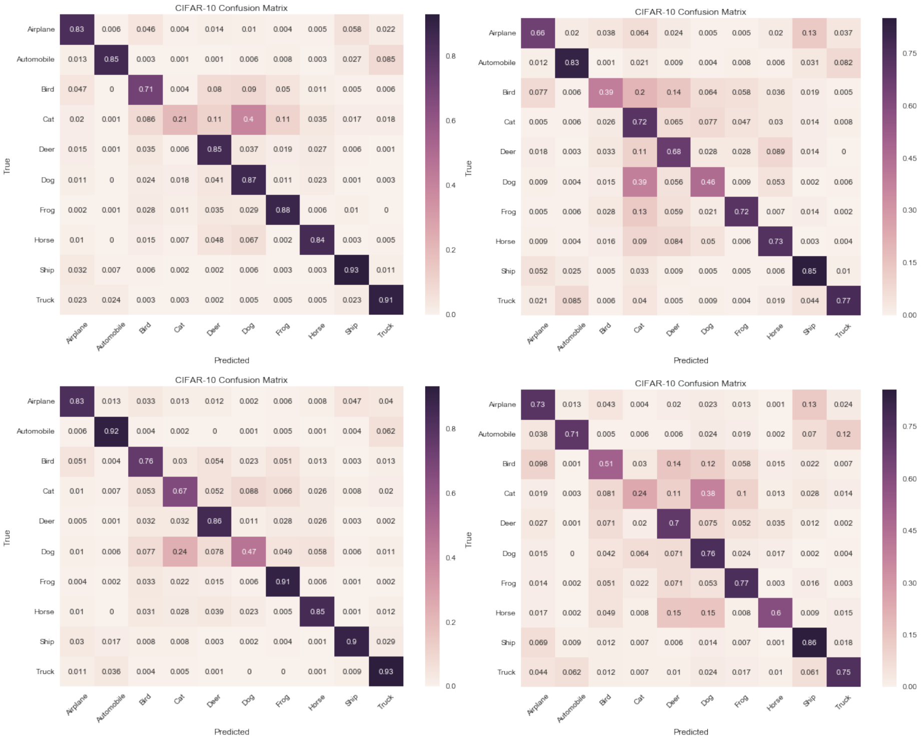


Fig. 6. Top left and right: cat under-sampling / over-sampling with perturbation. Bottom left and right: dog under-sampling / over-sampling with perturbation

under-sampled/over-sampled network with the gradient of the balanced network did not show to be as effective as using the same model's gradient. The average accuracy of an under-sampled model attack with adversaries generated from a different network was 53.8% while the same metric was 25.8% for the same model attack. Even that our training samples are within the same data domain, there are still huge differences on the gradients learned from the network.

4.3 Overlapping distributions

When classes in the dataset already have distributions that are very similar to one another the effects of adversaries seems to be even stronger. Figure 1 shows that for the pairs cat/dog and automobile/truck, misclassification naturally happens towards one another due to similarities in their feature space. On this case, our experiment shows that the adversarial attacks intensifies this phenomenon by increasing the number of times on which one class is picked over another. Figure 6 shows that both cat and dogs are increasingly misclassified between themselves when under-sampled datasets on both classes are used. While on the cat under-sampling case 40% of the samples were misclassified as dogs, on the oversampling 39% of the dogs were misclassified as cats. This results gives interesting insights, as it shows that the gradient sign is navigating around the target class distribution and when an overlap occurs it becomes easier to create an adversarial to the class with higher similarities.

5 Conclusion

This work sheds an important light on machine learning methods. Several real-life models are deeply concerned with possible vulnerabilities of their models, and studies on this field were being done for the past 20 years. Still, the imbalance learning problem remains one of the big questions in machine learning. Adversarial attacks are one more concern for those working with such systems, but they could also be seen as a tool that could forcibly make a model to stretch its occupation over the data domain space when we add adversaries to our training set [11]. Every new threat to predictive techniques is also a new improvement over model generalisation capabilities as the creation of the former could be used to improve the latter. Imbalanced datasets can be seen as one way of understanding adversarial attacks, as the lack of knowledge on a specific label seems to increase models vulnerabilities to such attacks

6 Future Work

Deep Neural Nets are seen by most people as black-box models as it is really hard to reason about what the model is actually learning. Tools like adversarial methods helps to extract insights from such methods and should be explored further. This work has performed tests on a dataset with a small number of classes, and in the future, datasets like ImageNet could be used to confirm our insights or discover new ones. The excessive amount of parameters in DNNs do not shield them from one of the most common effects in machine learning models - the domain shift caused by unseen data points. We strongly believe that the studies of adversaries could greatly help the understanding of the boundaries of such models.

References

1. Bengio, I.G.Y., Courville, A.: Deep learning (2016), http://www.deeplearningbook.org, book in preparation for MIT Press
2. Bengio, Y.: Rmsprop and equilibrated adaptive learning rates for non-convex optimization
3. Billovits, C., Eric, M., Agarwala, N.: Hitting depth: Investigating robustness to adversarial examples in deep convolutional neural networks (2016)
4. Canziani, A., Paszke, A., Culurciello, E.: An analysis of deep neural network models for practical applications. arXiv preprint arXiv:1605.07678 (2016)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. pp. 248–255. IEEE (2009)
6. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
7. Japkowicz, N., Stephen, S.: The class imbalance problem: A systematic study. Intelligent data analysis 6(5), 429–449 (2002)
8. Karpathy, A.: Convolutional neural networks for visual recognition (2016), https://cs231n.github.io/
9. Krawczyk, B.: Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence 5(4), 221–232 (2016)
10. Krizhevsky, A.: Cifar-10 and cifar-100 datasets (2009), https://www.cs.toronto.edu/~kriz/cifar.html
11. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial examples in the physical world. Tech. rep., arXiv (2016), http://arxiv.org/abs/1607.02533
12. Murphey, YI L, G.H.: Neural learning from unbalanced data (2004)
13. Papernot, N., McDaniel, P., Goodfellow, I.: Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. arXiv preprint arXiv:1605.07277 (2016)
14. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Berkay Celik, Z., Swami, A.: Practical black-box attacks against deep learning systems using adversarial examples. arXiv preprint arXiv:1602.02697 (2016)
15. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
16. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
17. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Advances in neural information processing systems. pp. 3320–3328 (2014)