Effectiveness of adversarial examples on convolutional neural networks

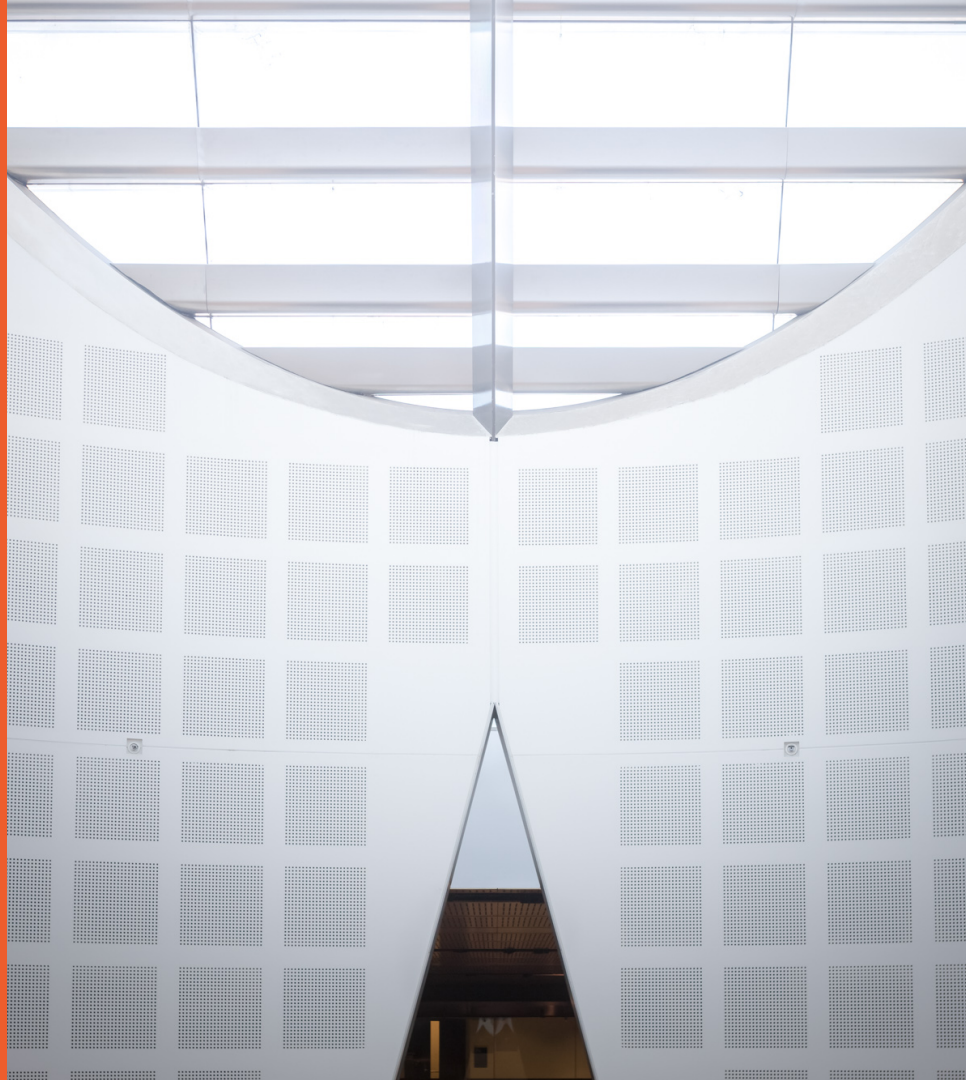
Presented by

Rafael Possas,

Research Engineer at Centre of Translational Data Science, University of Sydney

Master of Information Technologies (Research)

THE UNIVERSITY OF SYDNEY

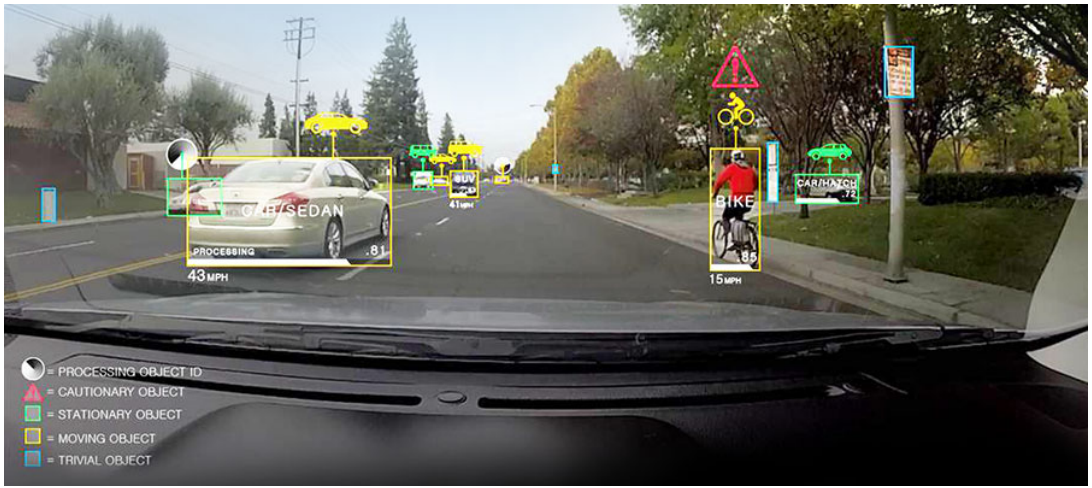


Contents

- Concepts
- Motivation
- Research Question
- Contributions
- Experiment design
- Results
- Conclusion

What is Computer vision?

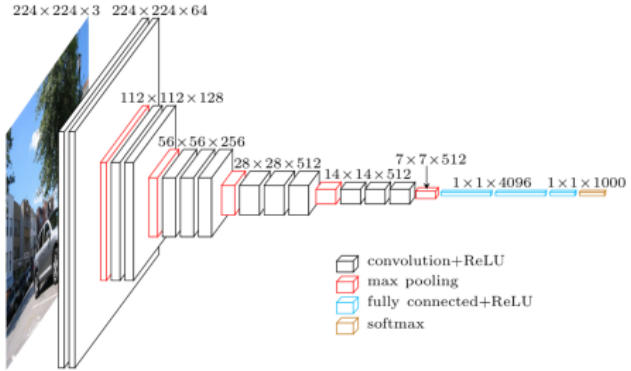
- *" Computer vision is an interdisciplinary field that deals with how computers can be made to gain high-level understanding from digital images or videos"* - Wikipedia



Example of object segmentation on a self-driving car [1]

Convolutional neural networks

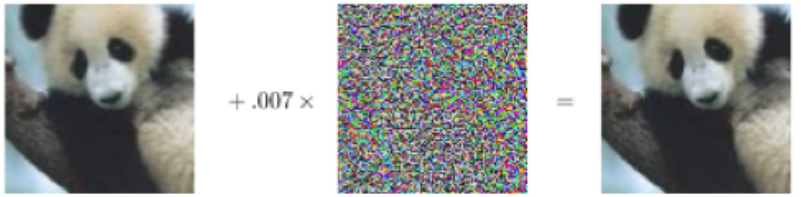
- "Type of feed-forward neural network which the connectivity pattern is inspired by the organization of the animal visual cortex" [2]
- It provides state of the art performance on data with spatial structure [1].



An example of CNN architecture, VGGNet [3]

Adversarial inputs

- Intentional modification on inputs to deliberately yield erroneous model outputs, while appearing unmodified to human observers.
- Depend on model information to successfully generate an adversary



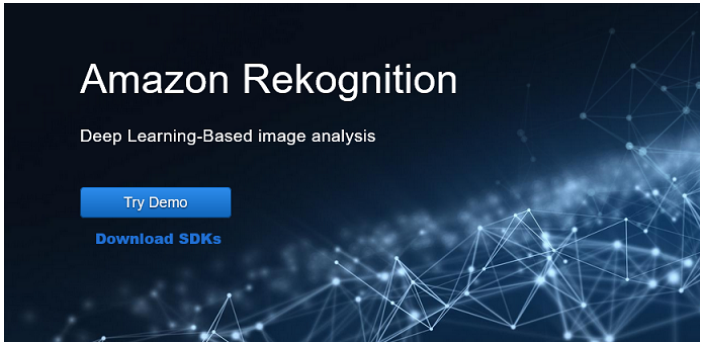
Adversarial crafting using Gradient Sign from Goodfellow et al. (2014) [3]

Imbalanced learning and domain shift

- **Imbalanced Learning**: When the underlying training set contains high variance on the number of samples for each class.
- **Domain Shift:** When the joint distributions of inputs and outputs differs between training and test stages.

Motivation

- Image recognition systems have been used widely in commercial applications
- Adversarial attacks are proven to work on state of the art convolutional neural networks (CNNs) architectures.
- Real world data is often imbalanced with lots of missing values
- There is no empirical evidence of adversarial attacks on class-imbalanced CNNs



Research question

- Would adversarial attacks be more effective on networks trained on Imbalanced datasets?
  - Does the attacks heavily depend on the targeted model internal knowledge?
  - Does classes with similar and different features space behave equally to adversarial attacks?

Contributions

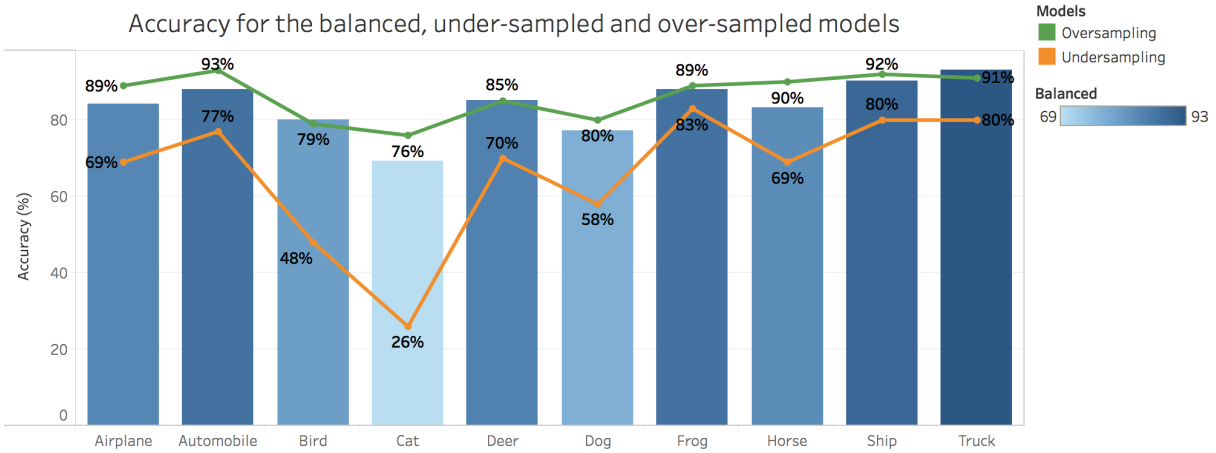
- Evaluation of the resilience of imbalanced CNNs to adversarial attacks using the white-box and black-box approach
- Investigation of classes with overlapping distributions and their relationship to both adversarial attacks and imbalanced learning problem

Experiment design (1): CNN Architecture

- First architecture to show that the depth of a network its critical to its performance
- Homogeneous architecture: Only 3x3 convolutions and 2x2 pooling from beginning to end
- We've changed the architecture to cope with the smaller input size (32x32).
- **Change**: Two 4096 FC => single 512 FC layer
- **Change**: 3 Convolution/pooling blocks instead of 4.

Experiment design (2): Imbalanced dataset

- **CIFAR-10**, 32x32 images on 10 non-hierarchical classes, 60,000 samples
- **Modified VGG**: achieved 83% overall accuracy on the balanced dataset
- **Baseline**: Balanced dataset vulnerability to adversaries
- **Tests**: 20 different imbalanced variations with under-sample and over-sample on each of the 10 classes



Experiment design (3): Gradient sign method

$$C(x+\delta) \approx C(x) + \epsilon * sign(\nabla C)$$
$$C(x+\delta) \approx C(x) - \epsilon * sign(\nabla C)$$

- Uses the gradient of any chosen class to add small noise to image pixels
- Only the gradient direction is used due to the ***sign*** operation, the magnitude of the perturbation is controlled by the **epsilon**
- Fast gradient sign x Iterative gradient sign
- Ascent x Descent perturbations
- **Epsilon 0.01**: Best trade-off between misclassification and image noise
- **Methods**: Fast gradient sign and ascent perturbations
- **Backpropagated gradient**: True class
- **Intuition**: Increase the cost function of the true class by the **epsilon** value in all gradient directions

Experiment design (4): White/Black-box attacks

- **Same model gradient (White-box):** Uses the underlying model knowledge to create optimal perturbations
- **Different model gradient (Black-box):** Uses an approximation of the true model, in this experiment, represented by the balanced network.
- Same data domain but datasets vary on the total number of samples.

Experiment workflow

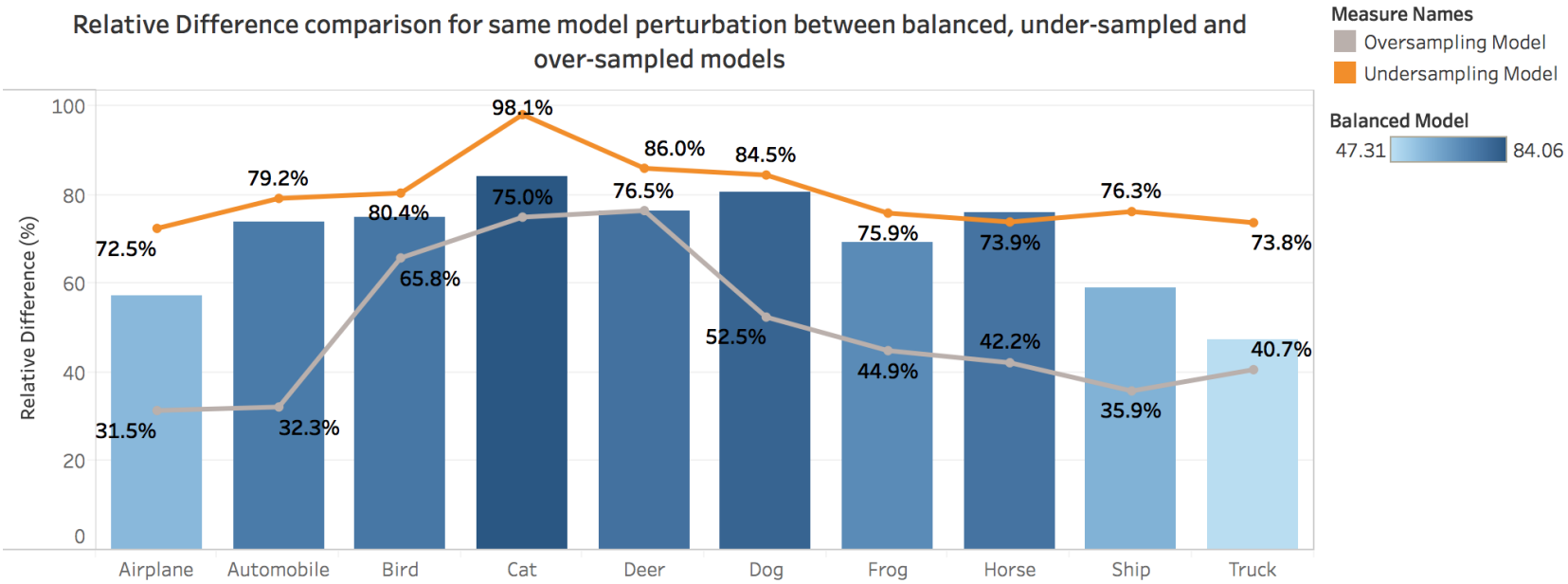
Results (1): Overall accuracy

- While the over-sampled models were more robust to adversaries, the under-sampled models had lower overall accuracy
- White-box attacks were generally more efficient than black-box attacks

	Black-box		White-box		
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%

Results (2): Relative difference on white-box attacks

- We calculated the percentage difference between the non-perturbed and perturbed model – the graph shows the results for the 3 variations on the white-box attack



Results (3): Overlapping distributions

Classes with similar features are often misclassified as one another

Stronger Classes

Balanced dataset with white-box attack

	Airplane	Automobile	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck
Dog	0.024	0.006	0.18	0.38	0.086	0.15	0.068	0.059	0.013	0.032

Dog under-sampling with white-box attack

	Airplane	Automobile	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck
Dog	0.025	0.01	0.17	0.32	0.12	0.09	0.12	0.1	0.01	0.033

Balanced dataset with white-box attack

	Airplane	Automobile	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck
Automobile	0.025	0.23	0.013	0.013	0	0.004	0.015	0.001	0.03	0.67

Automobile under-sampling with white-box attack

	Airplane	Automobile	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck
Automobile	0.068	0.16	0.01	0.014	0.016	0.007	0.026	0.016	0.18	0.51

Weaker Classes

Balanced dataset with white-box attack

	Airplane	Automobile	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck
Cat	0.035	0.005	0.15	0.11	0.12	0.31	0.14	0.045	0.026	0.06

Cat under-sampling with white-box attack

	Airplane	Automobile	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck
Cat	0.038	0.006	0.14	0.005	0.15	0.39	0.13	0.085	0.038	0.031

Balanced dataset with white-box attack

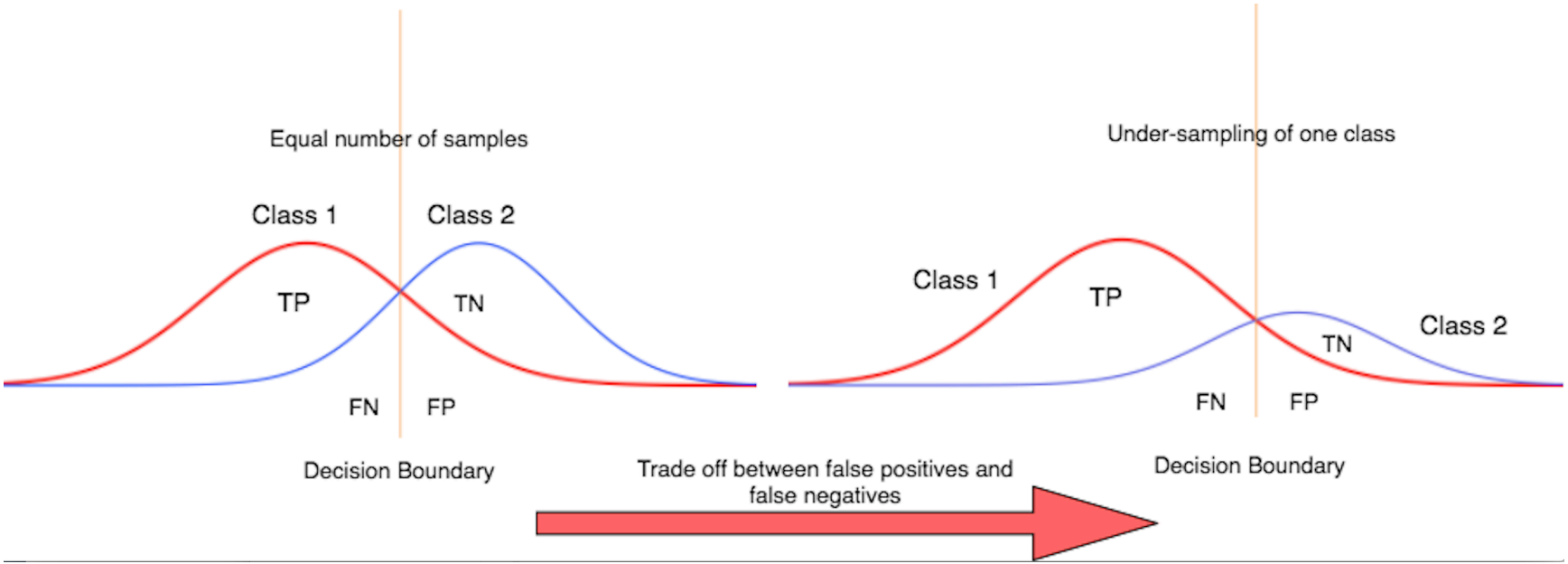
	Airplane	Automobile	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck
Truck	0.084	0.23	0.032	0.03	0.018	0.019	0.012	0.015	0.065	0.49

Truck under-sampling with white-box attack

	Airplane	Automobile	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck
Truck	0.11	0.42	0.022	0.043	0.038	0.01	0.023	0.039	0.089	0.21

Results (4): Decision boundary bias

Data imbalance causes a natural trade-off between over-sampled and under-sampled labels



Conclusions

- Under-sampled networks are more vulnerable to adversarial attacks while over-sampling increase robustness
- Black-box attacks are less effective than white-box which shows that gradients learned differs greatly even if we use the same data domain
- On the white-box attack with under-sampled training set, the non-dominant class was misclassified more often as the dominant class (overlapping distributions).
- CNNs shows linear behavior in high-dimensional spaces

Future Work?

- Use datasets with more classes
- Try different gradient sign methods with different combinations of epsilon values
- Do adversarial attacks are less or more vulnerable on different CNNs architectures?

Questions?