Adversarial Image Generation based on Various Neuron Coverage

Team8

20170181 Taeyoung Kim
20180650 Hyunjoon Cho
20205424 Sunjae Kwon

Table of Contents

Recap

- **DeepXplore & DLFuzz**

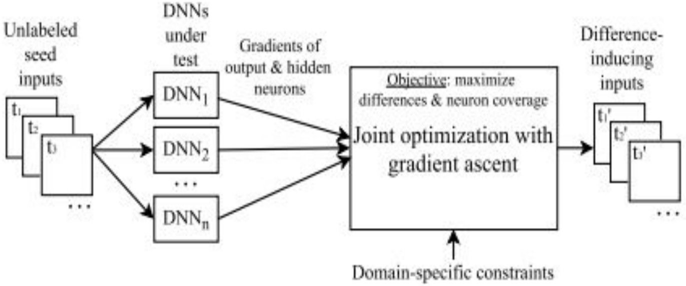


Figure 5: DeepXplore workflow.

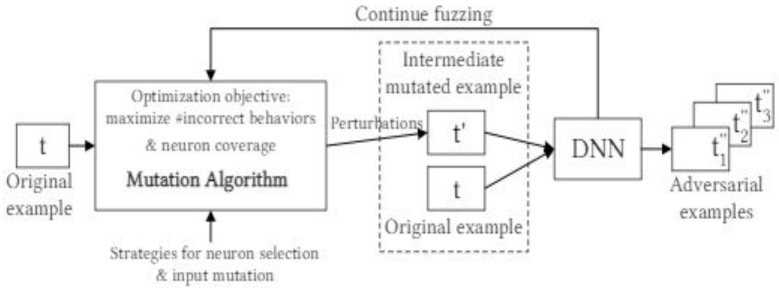


Figure 2: Architecture of DLFuzz

Pei, Kexin, et al. "Deepxplore: Automated whitebox testing of deep learning systems." proceedings of the 26th Symposium on Operating Systems Principles. 2017.

Guo, Jianmin, et al. "DLFuzz: differential fuzzing testing of deep learning systems." Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. 2018.

Recap

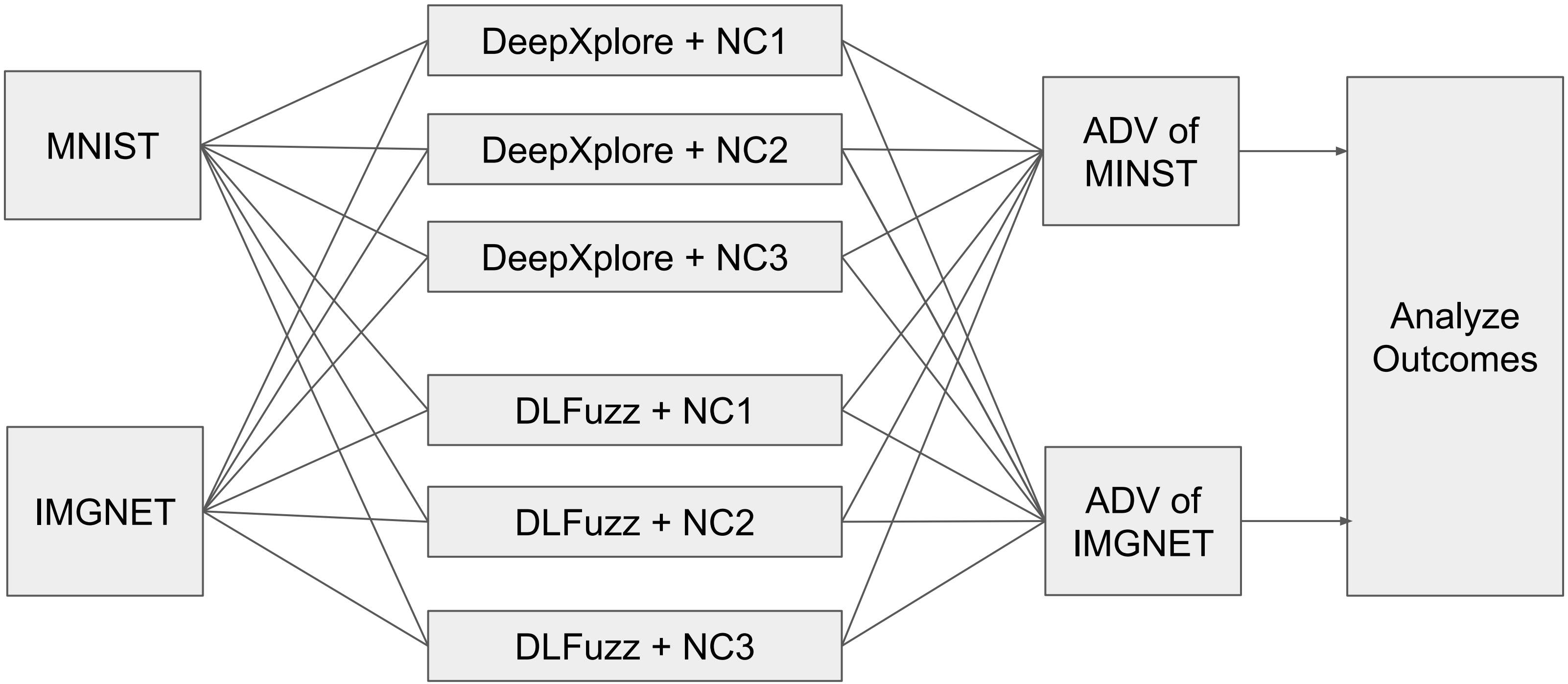
- Neuron Coverage (NC1)
  - Basic coverage, used in DeepXplore and DLFuzz
- k-multisection Neuron Coverage (NC2) Expectation: **Not Effective**
  - Section is bounded by low/high output from training
  - Sections are already covered by training data, less probable to show new behavior
- Strong Neuron Boundary Coverage (NC3) Expectation: **Effective**
  - Increased coverage may invoke more logic, resulting in unexpected behavior
  - Neurons are activated by output over threshold, thus upper bound would be more influential

Recap(Goal)

- Problem Statement
  - **DeepXplore** and **DLFuzz** depends on basic neuron coverage
  - In the meantime, various neuron coverage metric have been proposed

    ex) k-multisection Neuron Coverage, Strong Neuron Boundary Coverage.
  - Need for considering these various neuron coverage
- Project Goal
  - Find which coverage works **best** in creation of adversarial input

Method

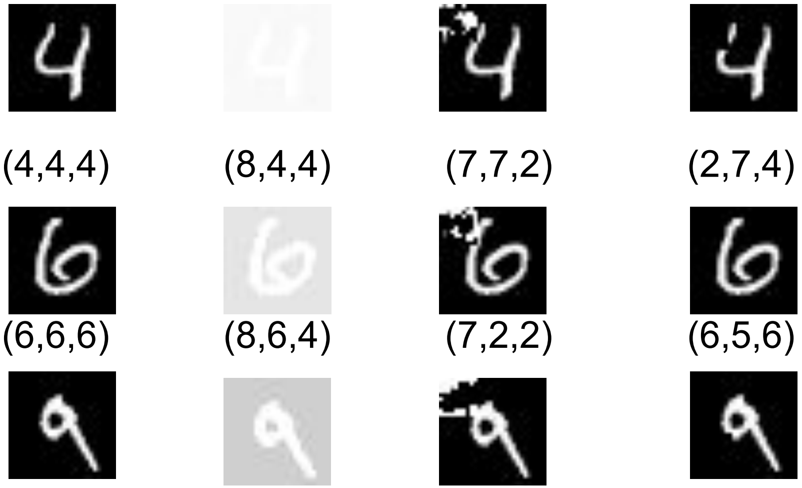

MNIST
IMGNET
DeepXplore + NC1
DeepXplore + NC2
DeepXplore + NC3
DLFuzz + NC1
DLFuzz + NC2
DLFuzz + NC3
ADV of MINST
ADV of IMGNET
Analyze Outcomes

Experimental set-ups

- **Dataset** : ImageNet / MNIST
- **Model** : Pretrained VGG / LeNet
- **Min-Max Calculation** : Extract from training set
    - IMGNET : Divide seeds into training set and adversarial generation set
    - MNIST : Use existing training set
- **Hyperparameters**
    - Loss coeff : adversarial loss 1, neuron activation loss 0.1
    - Grad coeff : learning rate 10, 20 steps, 100 seeds
    - Coverage : Threshold 0.2, 5 sections
- **Evaluation Metrics**
    1. Number of adversarial inputs generated
    2. Average time for generating single adversarial input
    3. Coverage of adversarial examples
    4. L2 distance between the original image and the adversarial image

Experimental Results

- **DeepXplore on MNIST**



	Original	Light	Occlusion	Blackout
NC1 (0.2 threshold)	(4,4,4)	(8,4,4)	(7,7,2)	(2,7,4)
NC2 (5-multisection)	(6,6,6)	(8,6,4)	(7,2,2)	(6,5,6)
NC3 (Strong)	(9,9,9)	(3,9,8)	(7,2,2)	(4,9,4)

Experimental Results

- **Results of DeepXplore on ImageNet**

***Each column corresponds to blackout, light, occlusion.**

	# Adv / # Initial			time per Adv (s)			Coverage			Avg L2		
NC1 (Neuron)	97/97	86/95	81/99	4.67	5.13	5.27	0.47	0.47	0.42	376	2586	492
NC2 (k-multi)	97/97	80/97	74/99	4.96	5.09	5.24	0.31	0.28	0.28	386	2708	487
NC3 (Strong)	97/97	76/97	84/96	4.70	4.64	4.73	0.48	0.27	0.44	381	2712	483

Experimental Results

- **DeepXplore on ImageNet**

* Constraint - Blackout does not work for ImageNet

NC1: Basic

Cassette

Projector

Table Lamp

Lampshade

NC2 : K-multi

Flute

Max Iteration:
20 times

Constraint:
Light

Oboe

Toilet Seat

Max Iteration:
20 times

Constraint:
OCCL

Bobsled

NC3 : Boundary

Tarantula

Coral Reef

Flamingo

Crane

Experimental Results

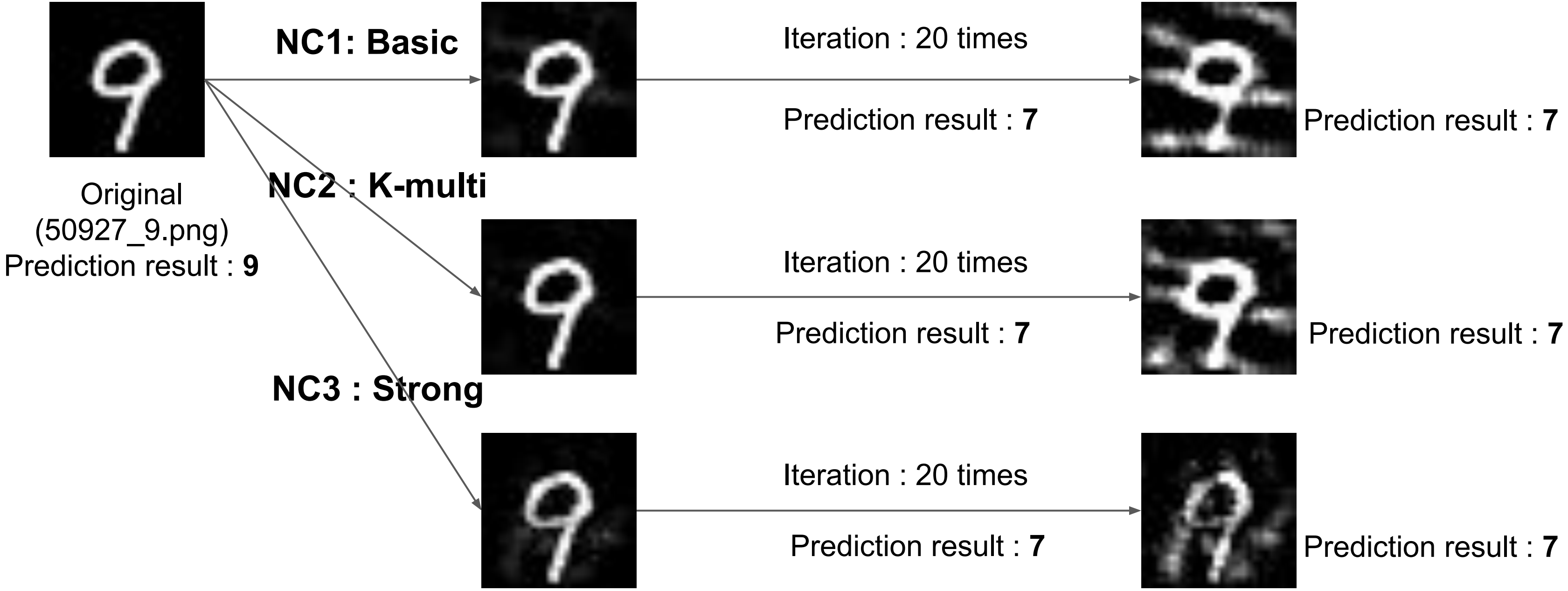
- **Results of DeepXplore on ImageNet**

*__Light__ constraint on **Left col**, **OCCL** constraint on **Right** col
*Run for **100 seeds** for each NC / constraint on colab w/ GPU

	# Adv / # Identical		time per Adv		Coverage		Avg L2	
NC1(Basic)	40 / 64	46 / 59	7.55	7.06	0.0736	0.0730	46657.4	13792.0
NC2(k-multi)	41 / 65	42 / 67	7.76	8.41	0.0753	0.0737	47316.8	13784.7
NC3(Boundary)	34 / 63	41 / 60	9.28	7.44	0.0754	0.0751	45248.1	14019.5

Experimental results

- **DLFuzz on MNIST**



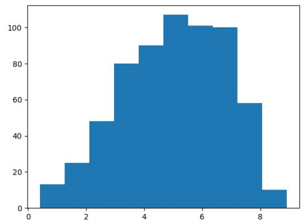
Experimental results

- Results of DLFuzz on MNIST

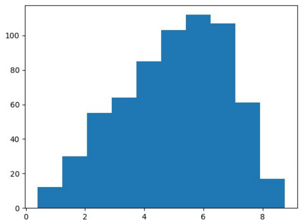
	# Adv	time per Adv	Coverage	# Seed	Avg L2 distance
NC1 (Basic)	557	6.18	0.67	37	**4.74**
NC2 (K-Multi)	632	**6.09**	**0.92**	42	5.01
NC3 (Strong)	**646**	6.12	0.48	**44**	5

- L2 distance distribution of # adv

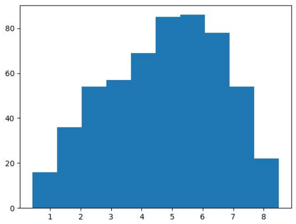
NC1



NC2

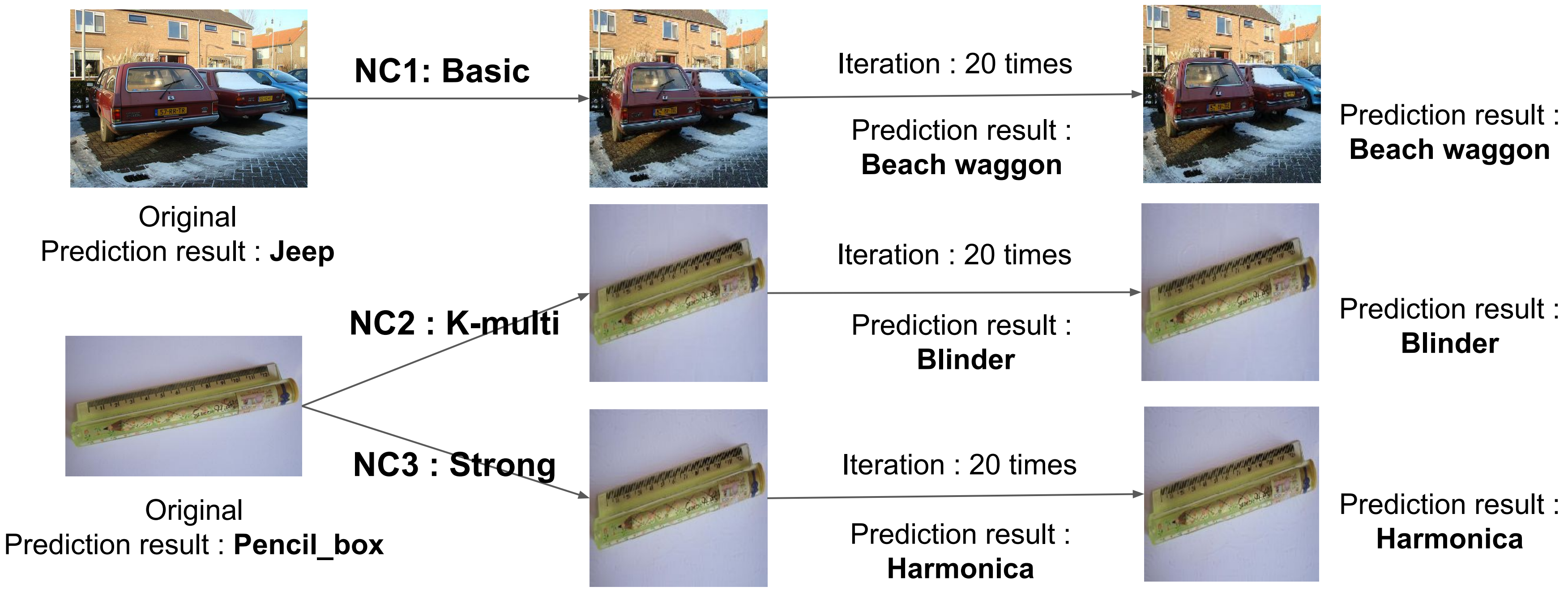


NC3



Experimental results

- **DLFuzz on ImageNet**



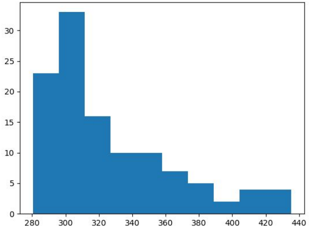
Experimental results

- **Results of DLFuzz on ImageNet**

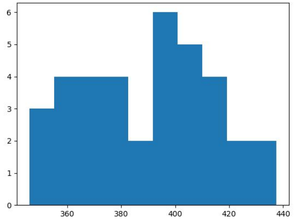
	# Adv	time per a Adv	Coverage	# Seed	Avg L2 distance
NC1 (Basic)	**178**	213	**0.496**	4	327.66
NC2 (K-Multi)	36	317	0.369	3	389.7
NC3 (Strong)	114	**199**	0.002	**5**	**326.3**

- **L2 distance distribution of # adv**

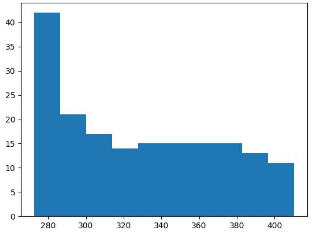
NC1



NC2



NC3



Analysis

- **Qualitative**
  - Different neuron coverage concepts enable each framework to find different adversarial images.
  - Constraints of generated adversarial images can influence the performance of the neuron coverage
- **Quantitative**
  - Considering evaluation metrics, 3 NCs have similar distribution on both frameworks.
  - No certain neuron coverage concept has better performance than others on both frameworks.

Future Works

- **Implementation**
  - Currently, boundary neuron coverage implementation needs to be fixed.
- **Improving frameworks**
  - Neuron loss only considers increasing output, need to be redefined so that it can decrease its output to hit lower boundary.
  - Optimize parameters; step size, weight of differential behavior and NC
- **Analysis**
  - Each coverage showed various generated inputs, but we don’t know they are diverse, or have some pattern.
  - By measuring coverage without adversarial example, check whether adversarial examples really increases coverage.

Conclusions

- We confirmed the performance of 3 different neuron coverage concepts on 2 different DNN testing tools.
- No certain neuron coverage concept has better performance than others on both frameworks.
- However, constraints on generated images and type of neuron coverage enable each framework to generate different adversarial images.
- Thus, further study is required about new boundary concept and loss function to aggregate various neuron coverage.