ICANN19

DCT:Differential Combination Testing of Deep Learning Systems

Chunyan Wang, Weimin Ge, Xiaohong Li*, Zhiyong Feng
College of Intelligence and Computing, Tianjin University, Tianjin, China
chunyan_wang@tju.edu.cn, xiaohongli@tju.edu.cn

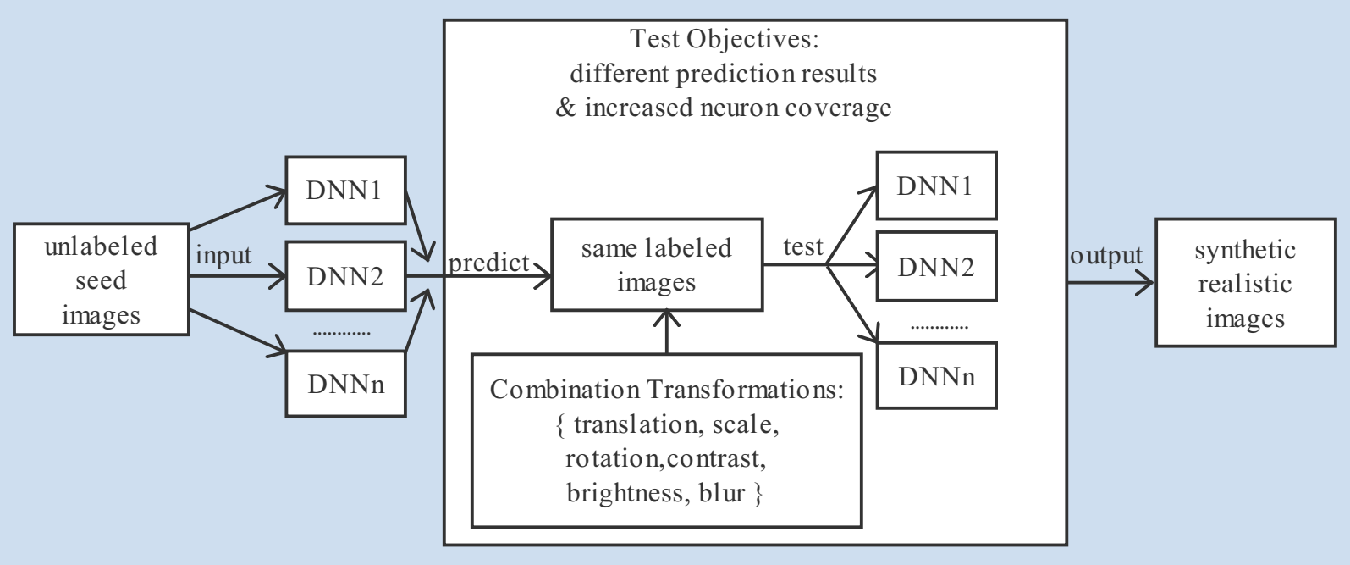
Introduction

- Deep learning (DL) systems are increasingly used in security-related fields. However the DL models are difficult to test and existing DL testing relies heavily on manually labeled data and often fails to expose erroneous behavior for corner inputs.
- In this paper, we propose Differential Combination Testing(DCT), an automated DL testing tool for systematically detecting the erroneous behavior of more corner cases without relying on manually labeled input data or manually checking the correctness of the output behavior.

Method

We take an overview of DCT and then describe three key points of DCT - differential testing, neuron coverage and image combination transformations.

1. Overview of DCT:


unlabeled seed images
input
DNN1
DNN2
....
DNNn
predict
Test Objectives:
different prediction results
& increased neuron coverage
same labeled images
test
DNN1
DNN2
....
DNNn
Combination Transformations:
{ translation, scale, rotation,contrast, brightness, blur }
output
synthetic realistic images

The workflow of DCT, which leverages combination transformations to generate synthetic realistic images with neuron coverage and prediction feedback as guidance.

2. We apply differential testing, that is, we use multiple DNNs with similar functions as cross-references, so that the test inputs are not necessarily labeled images and DCT can automatically check the correctness of output behaviors.

3. We use neuron coverage proposed by Pei et al. [1] to measure the test degree of the internal logic of the DNN.

$$NCoverage = \frac{|\{n|\forall x \in I, \phi(n,x) > t\}|}{|N|} \quad (1)$$

where $N = \{n_1, n_2,...\}$ represents all neurons, $I = \{x_1, x_2,...\}$ represents all test inputs, $\phi(n,x)$ is a function that returns the output value of a neuron $n \in N$ for a given test input $x \in I$ and t represents the threshold.

4. An image x' is generated after a transformation tr and a corresponding parameter p on x (denoted as $x \xrightarrow{(tr,p)} x'$).An image x' is generated after a sequence of transformations($x \xrightarrow{(tr_1,p_1)} x_1$,$x_1 \xrightarrow{(tr_2,p_2)} x_2$,......,$x_{n-1} \xrightarrow{(tr_n,p_n)} x'$) (denoted as $x \xrightarrow{(tr_1,p_1),(tr_2,p_2),...,(tr_n,p_n)} x'$).

Conclusion

DCT can automatically test the reliability and accuracy of DL systems. DCT generates corner test images by applying different combination transformations on seed images and guided by differential testing and neuron coverage. The tool does not need to rely on manual labeled data and can automatically identify erroneous behaviors.

References

[1] Kexin Pei, Yinzhi Cao, Junfeng Yang, and Suman Jana. Deepxplore: Automated whitebox testing of deep learning systems. In *proceedings of the 26th Symposium on Operating Systems Principles*, pages 1–18. ACM, 2017. doi: 10.1145/3132747.3132785.

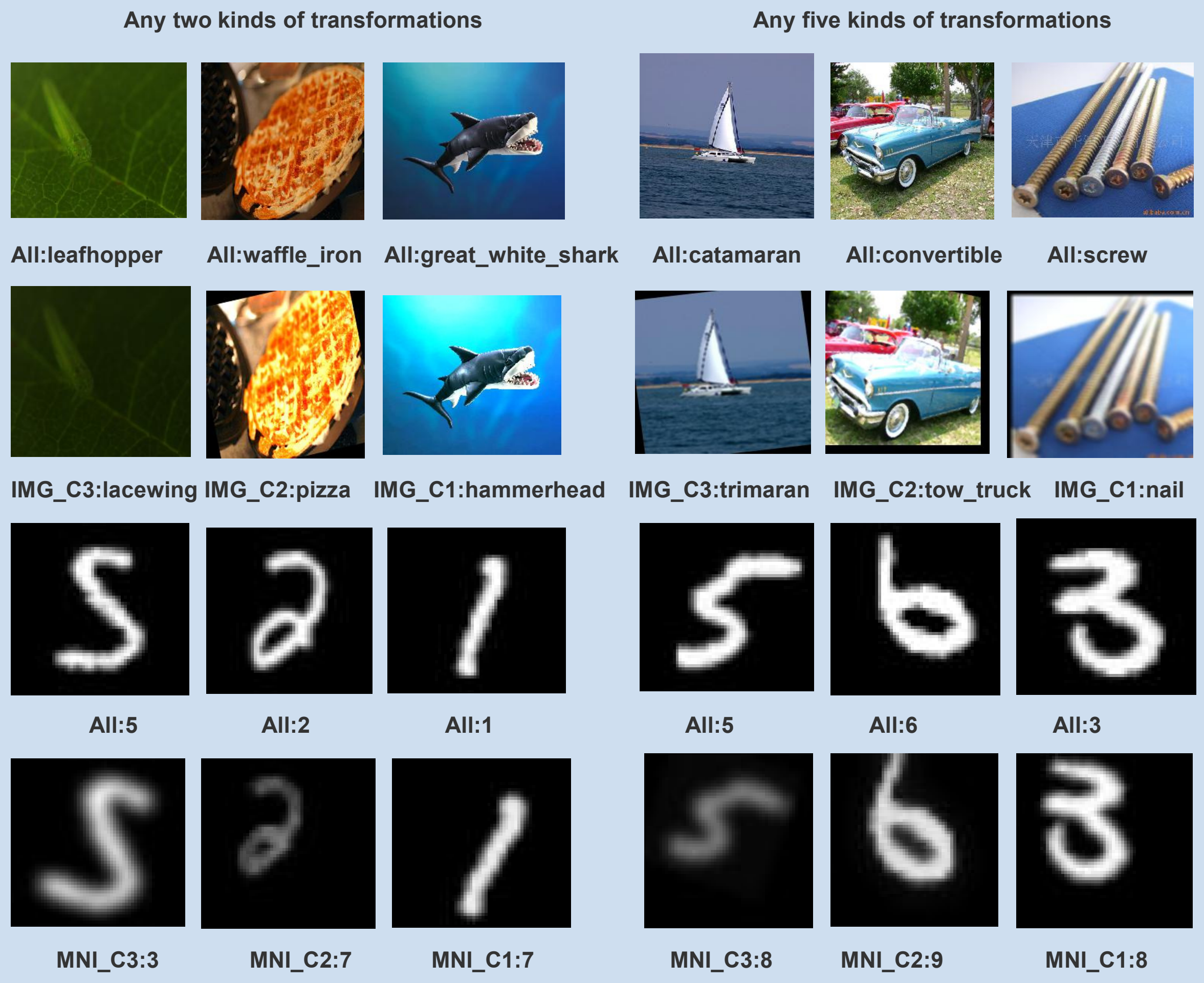
[2] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Experiments

We select two popular public datasets, MNIST and ImageNet, as the evaluation datasets. For each dataset, we study DCT on three popular DNN models. We provide a summary of the two datasets and corresponding DNNs in the paper. We use six different types of image transformations. They are translation, scale, rotation, contrast adjustment, brightness adjustment and blur. We implement these transformations by using OpenCV.

Results

The Figure below shows some synthetic test cases and the corresponding erroneous behaviors generated by DCT for MNIST and ImageNet.

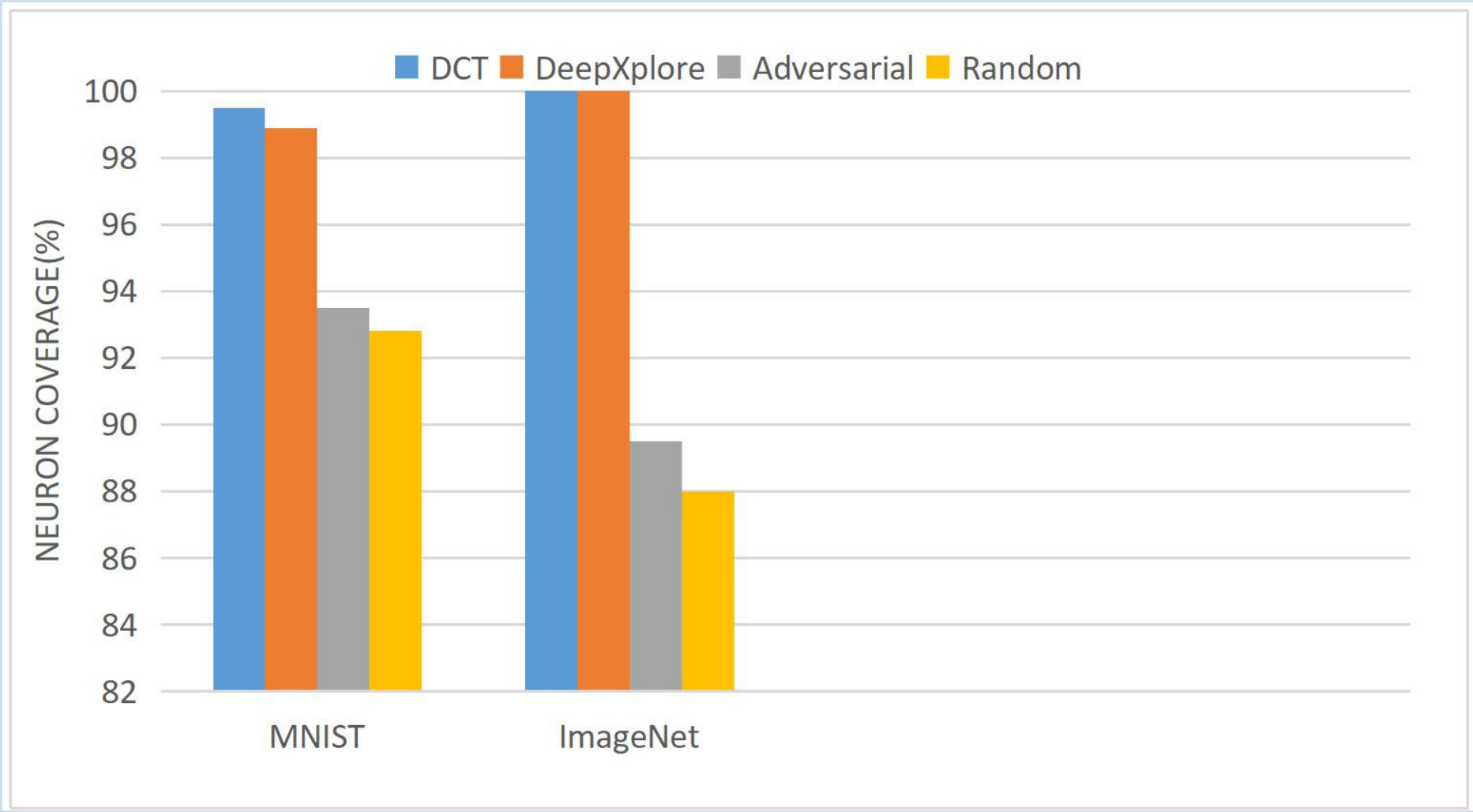


The table below shows that by randomly combining six different transforms, we get thousands of valid synthetic images that represent the erroneous behaviors of DNN models. It can be seen that as the kind of random image combination transformations increases, the synthetic images that detect the erroneous behaviors of the DNN models also increase significantly.

Variable k	DNN name					
	MNI_C1	MNI_C2	MNI_C3	IMG_C1	IMG_C2	IMG_C3
2	348	230	284	1017	987	870
3	695	541	556	1370	1358	1127
4	1073	741	845	1988	2035	1616
5	1476	1078	1308	2645	2716	2303

Performance Evaluation

We use two indicators to evaluate the performance of DCT: neuron coverage and execution time. We compare neuron coverage achieved by four different methods. They are: (1) DCT, (2) DeepXplore [1], (3) adversarial testing [2], and (4) random selection testing. The results are shown below.



We can see from the results that the DCT covers an average of 8.25% and 9.3% more neuron coverage than adversarial testing and random testing, while DCT and DeepXplore are comparable in neuron coverage and can be both used for more thorough testing on neurons. We further compared the total time taken by DCT and DeepXplore on MNIST and ImageNet. The results are shown in table below. As can be seen from the table, no matter what kind of image combination transformations the DCT runs, it is much shorter than the execution time of DeepXplore.

	DCT Variable k				DeepXplore
	k=2	k=3	k=4	k=5	
MNIST	2.18	5.5	10.8	21.9	36.7
ImageNet	270.8	361.7	509.2	720.4	943.71