DeepPath: Path-driven Testing Criteria for Deep Neural Networks

Dong Wang[1] Ziyuan Wang[2] Chunrong Fang[1] Yanshan Chen[2] Zhenyu Chen[1]

[1] Institute of Software, Nanjing University, Nanjing, China

[2] School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, China

***Abstract*—Inspired by path-oriented testing, we propose a series of path-driven testing criteria, called *DeepPath*, to comprehensively calculate coverage in deep neural networks (DNNs). Four DNN models and four adversarial attack techniques are used to evaluate the effectiveness of *DeepPath*. The experimental results illustrate that DeepPath are more discriminating to measure test adequacy of DNNs in practice, as well as more useful for recognizing adversarial attack test inputs.**

***Key Words*—deep neural networks, testing criteria, path coverage**

I. Introduction

Due to the wide-scale deployment of deep learning systems, testing should be an important way to ensure their safety. However, research on DNN testing is still very preliminary. Testing criteria is first valued in DNN testing in recent works [1] [2] [3] [4].

A deep neural network, which consists of neurons in multiple layers and connections among neurons in neighbor layers, can be considered as a weighted directed connected graph that consists of nodes and weighted directed edges. Previous peer work paid attention on nodes in graphs (or neurons in DNNs). Meanwhile, weighted directed edges in graphs (or connections in DNNs) which transmit information from pre-layer neurons to post-layer neurons with individual weights, also should be considered.

Furthermore, these connections form a huge number of paths from input layer neurons to output layer neurons. These paths represent the information transmission in an entire DNN model. The outputs are determined by the computational flows based on neurons and connections. It is well known that path-oriented testing is always used for high-dependable testing of traditional software. However, control flow and data flow paths have no direct meaning for functional requirements of DNNs.

Based on the above observations, we propose a series of testing criteria by focusing on paths in DNNs. We believe that it can open a new perspective and some emerging directions of white-box testing for DNNs. Our contributions could be summarized as follows: 1) Novel testing criteria are proposed to evaluate testing adequacy of DNNs in-depth, *DeepPath*, a series of scalable testing criteria based on paths that consist of sequentially linked connections. 2) We demonstrate that *DeepPath* performs better effectiveness than neuron criterion. The measure discrimination of testing criteria is studied on the original test data and adversarial attack test inputs.

II. *DeepPath*: Path-driven Coverage Criteria

Given N is a deep neuron network that contains an input layer, t hidden layers and an output layer. There are a_i neurons $n_{i,1}$, $n_{i,2}$, ..., n_{i,a_i} in the i-th hidden layer $L_i(0 \leq i \leq t-1)$. The output value of neuron $n_{i,k}$ under a test input x is $out(n_{i,k}, x)$. The connection from neuron n_{i,k_1} to n_{i+1,k_2} is $e_{i;k_1,k_2}$, where n_{i,k_1} and n_{i+1,k_2} are input neuron and output neuron of such a connection.

Definition 1: (**Neuron State**). The neuron state could be divided into activated or inactivated. For a test input x, the state of neuron $n_{i,k}$ is:

$$s_{i,k}(x) = \begin{cases} 1, out(n_{i,k}, x) \in (\beta, +\infty) \\ 0, out(n_{i,k}, x) \in (-\infty, \beta] \end{cases}$$

where β is a threshold parameter. Generally, we use the RELU activation function, usually the threshold $\beta = 0$. Whether a neuron is activated or not, represents whether calculation of the neuron contributes to final output of the entire DNN. Based on neuron state, the subpath state defines as follows.

Definition 2: (**Subpath State**). Given a l-length subpath $p_{i;k_0,k_1,...,k_l}$ that consists of $l+1$ neurons n_{i,k_0}, n_{i+1,k_1}, ..., n_{i+l,k_l}, the state of such a subpath $p_{i;k_0,k_1,...,k_l}$ of an input x:

$$s_{i;k_0,k_2,...,k_l}(x) = (s_{i,k_0}(x), s_{i+1,k_1}(x), ..., s_{i+l,k_l}(x))$$

where $s_{i,k_0}(x)$ is the state of neuron n_{i,k_0} of x and etc. When $l = t-1$, the subpath is full path. And the subpath equals to nueron when $l = 0$. For a test input x, the state of a given l-length subpath must be one of following 2^{l+1} subpath state schemas:

$$(0,0,...,0,0),\ (0,0,...,0,1),$$
$$(0,0,...,1,1),\ ...,\ (1,1,...,1,1)$$

Here we say it is covered by the test input x. By calculating the number and percent of covered connection state schemas, we can define DeepPath coverage criteria for DNNs.

Definition 3: (**Path Coverage**). For a test suite T, we define the Path coverage as:

$$PCov(T) = \sum_{i=1}^{t-l} \sum_{k_0=1}^{a_i} ... \sum_{k_l=1}^{a_{i+l}} \frac{|Cov_{i;k_0,...,k_l}(T) \bigcap CovReq_{i;k_0,...,k_l}|}{|CovReq_{i;k_0,...,k_l}|}$$

Different definition of set $CovReq_{i;k_0,...,k_l}$ for the path $p_{i;k_0,...,k_l}$ will lead to different concrete path coverage criteria. We define three concrete path coverage criteria following:

- **l-length Strong Activated Path Coverage (l-SAP).** A path $p_{i;k_0,...,k_l}$ is SAP-covered by test suite T, if test suite T covers all path state schemas of set $CovReq_{i;k_0,...,k_l}$:

$$CovReq_{i;k_0,...,k_l} = \{(1,1,\cdots,1)\}$$

978-1-7281-0492-8/19/$31.00 ©2019 IEEE
DOI 10.1109/AITest.2019.00013

- **l-length Output Activated Path Coverage (l-OAP).** Generally, in a path, if one of middle neurons is activated, then subsequent neurons continue to transmit activated state to last neuron.We define that, a path is in output activated state, if a neuron in the path is firstly in activated state, and all subsequent neurons is in activated state. OAP measures if a test suite covers all path state schemas of path in output activated state. A path $p_{i;k_0,\ldots,k_l}$ is OAP-covered by test suite T, if test suite T covers all path state schemas of set $CovReq_{i;k_0,\ldots,k_l}$:

$$\begin{aligned} CovReq_{i;k_0,\ldots,k_l} =& \{(s_{i,k_0}, s_{i+1,k_1}, \cdots, s_{i+l,k_l}) \\ & |s_{i,k_0}, \cdots, s_{i+j,k_j} = 0 \\ & \&\ s_{i+j+1,k_{j+1}}, \cdots, s_{i+l,k_l} = 1 \\ & \&\ j\ \in [0, l)\} \end{aligned}$$

For example, a path $(n_{1,1}, n_{2,2}, n_{3,3})$ has 3 valid path state schemas, $CovReq_{1;1,2,3}$ = (0, 0, 1), (0, 1, 1),(1, 1, 1). In l-OAP, a l-length path need to cover l path state schemas which make output neurons activated, because the first activated neuron can be anyone of these l neurons in the path. If we only consider one situation that the first neuron of a path is firstly activated, l-OAP is equivalent to l-SAP.

- **l-length Full State Path Coverage (l-FSP).** A path $p_{i;k_0,\ldots,k_l}$ is FSP-covered by test suite T, if test suite T covers all path state schemas of set $CovReq_{i;k_0,\ldots,k_l}$:

$$\begin{aligned} CovReq_{i;k_0,\ldots,k_l} =& \{(s_{i,k_0}, s_{i+1,k_1}, \cdots, s_{i+l,k_l}) \\ & \mid s_{i,k_0}, s_{i+1,k_1}, \cdots, s_{i+l,k_l} \in \{0, 1\}\} \end{aligned}$$

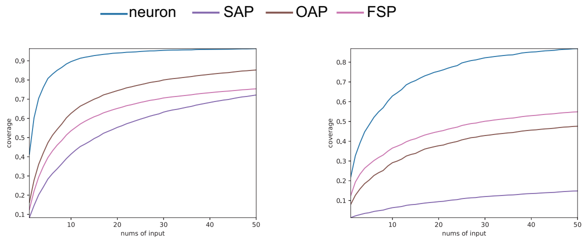
III. Experiment

Research Qestions: 1) Compared to neuron coverage, does *DeepPath* perform a better discrimnation? 2) Is *DeepPath* more effective for discriminating adversarial test samples?

Evaluation Subjects: We choose MNIST dataset as experimental target. As for models, we select a well-defined model Lenet-5 and three customized fully connected models, which separately include 3, 5 and 10 hidden layers (called L-3, L-5 and L-10 in the following). To obtain enough attack samples, we generate adversarial data by using some state-of-the-art adversarial attack techniques, including FGSM, JSMA and two decision based techniques, Gaussian Noise and Uniform Noise separately.

Evaluation Setup: We first trained the four studied models with training data. Then we generated four adversarial test dataset with attack technologies mentioned above through test data for each model. For RQ1, we randomly select 50 input data from all datasets of each studied model. Then we further calculated the accumulation of different coverage criteria to find out the difference in coverage growth. As for RQ2, we calculated coverage through the neuron and 2-length path criteria for each dataset to find out the comparative result.

Experimental Results: In the experiments, we choose 0 and 0.5 as threshold β to determine whether the neuron is activated or not. Fig. 1 illustrates the coverage growth trend diagram. And as shown in Table I, we exhibit the coverage of DeepPath(2-SAP, 2-OAP, 2-FSP) and neuron criterion.



(a) threshold=0

(b) threshold=0.5

Fig. 1: Coverage growth of various criteria

TABLE I: Coverage Results

Testing Criteria		DNN	threshold	origin	origin +FGSM	origin +JSMA	origin +Gaussian	origin +Uniform	all
Path Cov(%)	SAP Cov(%)	L-3	β=0	98.11	98.19	98.79	98.17	98.12	98.83
			β=0.5	91.85	92.19	92.57	92.13	92.14	92.81
		L-5	β=0	97.39	98.54	98.29	97.97	97.95	98.69
			β=0.5	88.37	89.26	89.55	88.63	88.65	89.92
		L-10	β=0	95.58	96.69	96.90	96.27	96.25	97.06
			β=0.5	57.04	59.41	62.45	57.77	57.73	62.92
		Lenet-5	β=0	98.85	99.20	99.35	98.92	98.90	99.45
			β=0.5	73.21	73.62	74.14	73.26	73.26	74.34
	OAP Cov(%)	L-3	β=0	98.47	98.54	99.02	98.50	98.48	99.04
			β=0.5	93.95	94.18	94.43	94.17	94.16	94.60
		L-5	β=0	97.78	98.74	98.50	98.23	98.22	98.84
			β=0.5	92.83	93.36	93.45	92.98	92.99	93.66
		L-10	β=0	96.50	97.30	97.43	96.98	96.95	97.54
			β=0.5	78.94	80.21	81.59	79.38	79.35	81.84
		Lenet-5	β=0	75.98	78.19	76.65	76.72	76.15	78.64
			β=0.5	86.57	87.64	88.37	86.73	86.78	88.74
	FSP Cov(%)	L-3	β=0	85.51	85.82	86.16	85.56	85.55	86.25
			β=0.5	84.24	84.46	84.62	84.39	84.38	84.74
		L-5	β=0	85.47	86.21	86.08	85.75	85.73	86.34
			β=0.5	83.45	83.88	83.93	83.56	83.57	84.09
		L-10	β=0	84.83	85.48	85.55	85.20	85.18	85.66
			β=0.5	76.98	77.76	78.50	77.29	77.27	78.63
		Lenet-5	β=0	61.50	63.87	62.08	62.23	61.65	64.29
			β=0.5	75.61	76.50	77.00	75.81	75.82	77.25
Neuron Cov(%)		L-3	β=0	100.00	100.00	100.00	100.00	100.00	100.00
			β=0.5	99.11	99.11	99.11	99.11	99.11	99.11
		L-5	β=0	99.42	100.00	100.00	99.71	99.71	100.00
			β=0.5	98.26	98.26	98.26	98.26	98.26	98.26
		L-10	β=0	98.16	98.41	98.53	98.41	98.28	98.53
			β=0.5	95.59	95.71	95.59	95.59	95.59	95.71
		Lenet-5	β=0	100.00	100.00	100.00	100.00	100.00	100.00
			β=0.5	97.35	97.35	97.35	97.35	97.35	97.35

Remarks: According to Fig. 1 and Table I, we summarize the results as follows. The coverage of all criteria generally come up to a relatively high percent at a relatively amount of test data. But *DeepPath* illustrates a more moderate speed of coverage growth than neuron coverage. Adversarial data could contribute to increase the coverage based on *DeepPath*. The effect would be more pronounced when the models are more complicated and the criteria are stricter.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (61832009, 61772259).

References

[1] K. Pei, Y. Cao, J. Yang, and S. Jana, "Deepxplore: Automated whitebox testing of deep learning systems," in *SOSP*, 2017, pp. 1–18.

[2] L. Ma, F. Juefei-Xu, F. Zhang, J. Sun, M. Xue, B. Li, C. Chen, T. Su, L. Li, Y. Liu *et al.*, "Deepgauge: multi-granularity testing criteria for deep learning systems," in *ASE 2018*.

[3] Y. Sun, X. Huang, and D. Kroening, "Testing deep neural networks," in *ASE 2018*.

[4] L. Ma, F. Zhang, M. Xue, B. Li, Y. Liu, J. Zhao, and Y. Wang, "Combinatorial testing for deep learning systems," *arXiv preprint arXiv:1806.07723*, 2018.