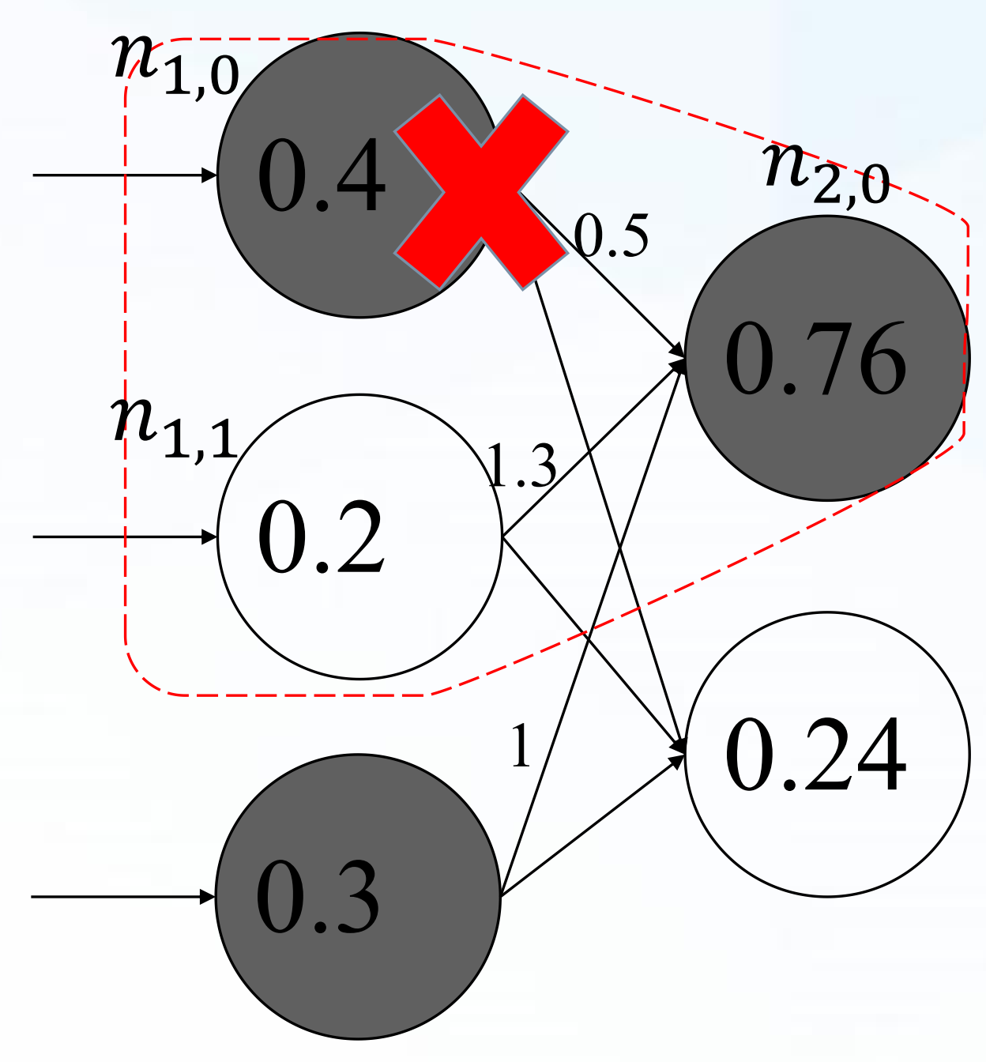
面向深度神经网络的贡献覆盖测试

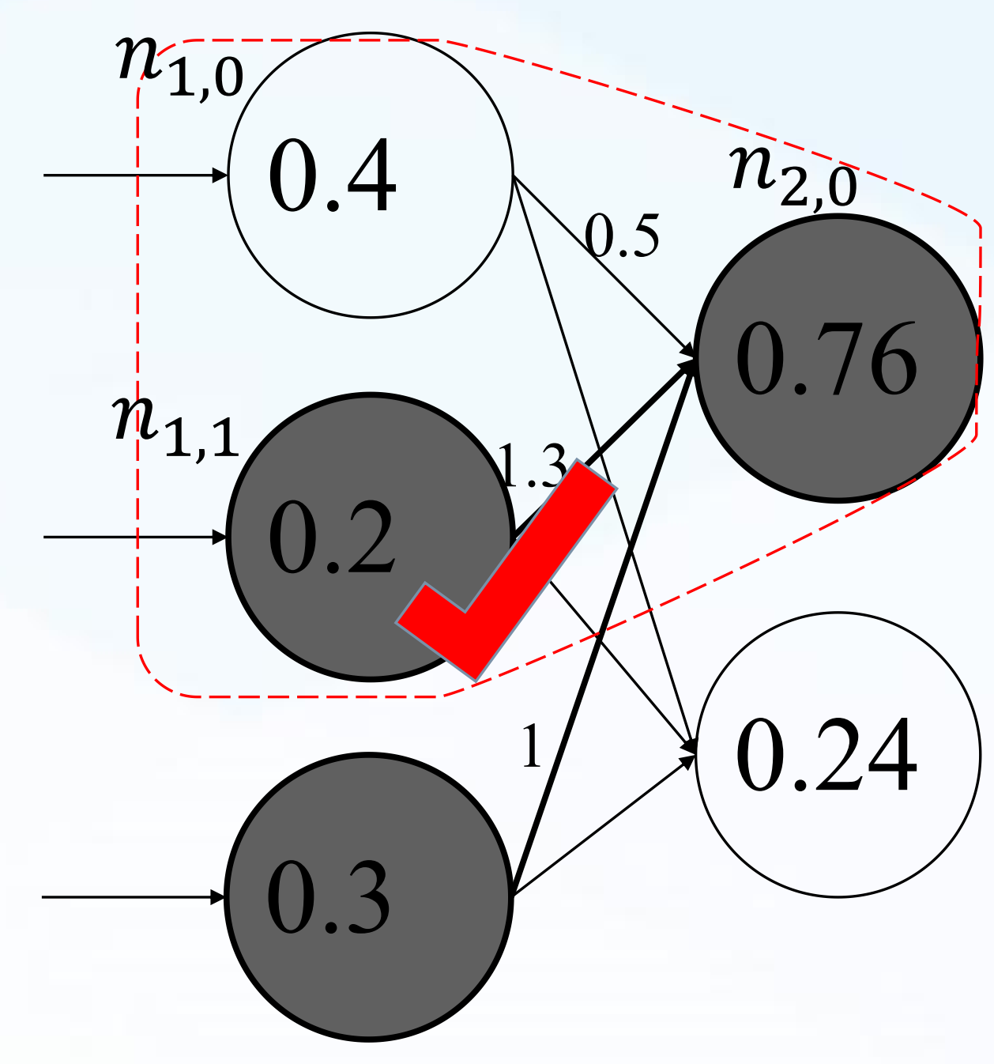
（DeepCon: Contribution Coverage Testing for Deep Learning Systems, SANER 2021）

周志阳，窦文生，刘杰，张晨昕，魏峻，叶丹
{zhouzhiyang18, wsdou, ljie}@otcaix.iscas.ac.cn

- □ 深度神经网络（DNN）广泛应用于安全敏感的领域，需被充分地测试
- □ DNN的内部数据逻辑存于模型的神经元、权重、层中，而非代码片段中，需要制定新的覆盖度量标准来反映测试的充分性
- □ 基于神经元输出的覆盖度量标准[1-3]的仅考虑到神经元而忽略了其射出的连接权重，对重要“逻辑”的提取存“misfitting”

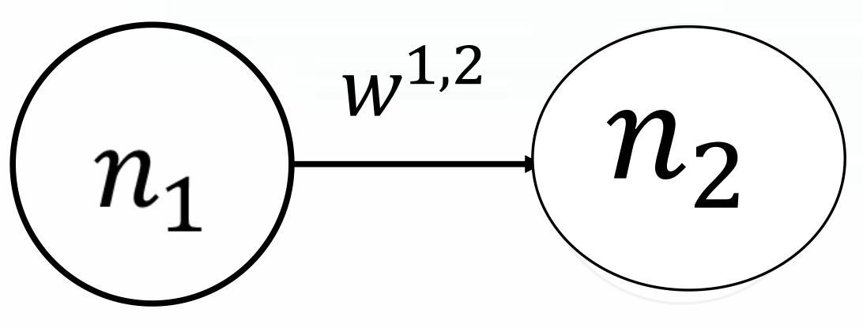


基于神经元输出的覆盖度量标准所提取的激活神经元（灰色节点）



由于0.4*0.5 < 0.2*1.3，则$n_{1,1}$对预测结果的做出的贡献比$n_{1,0}$更大

一、贡献覆盖率度量标准（DeepCon）

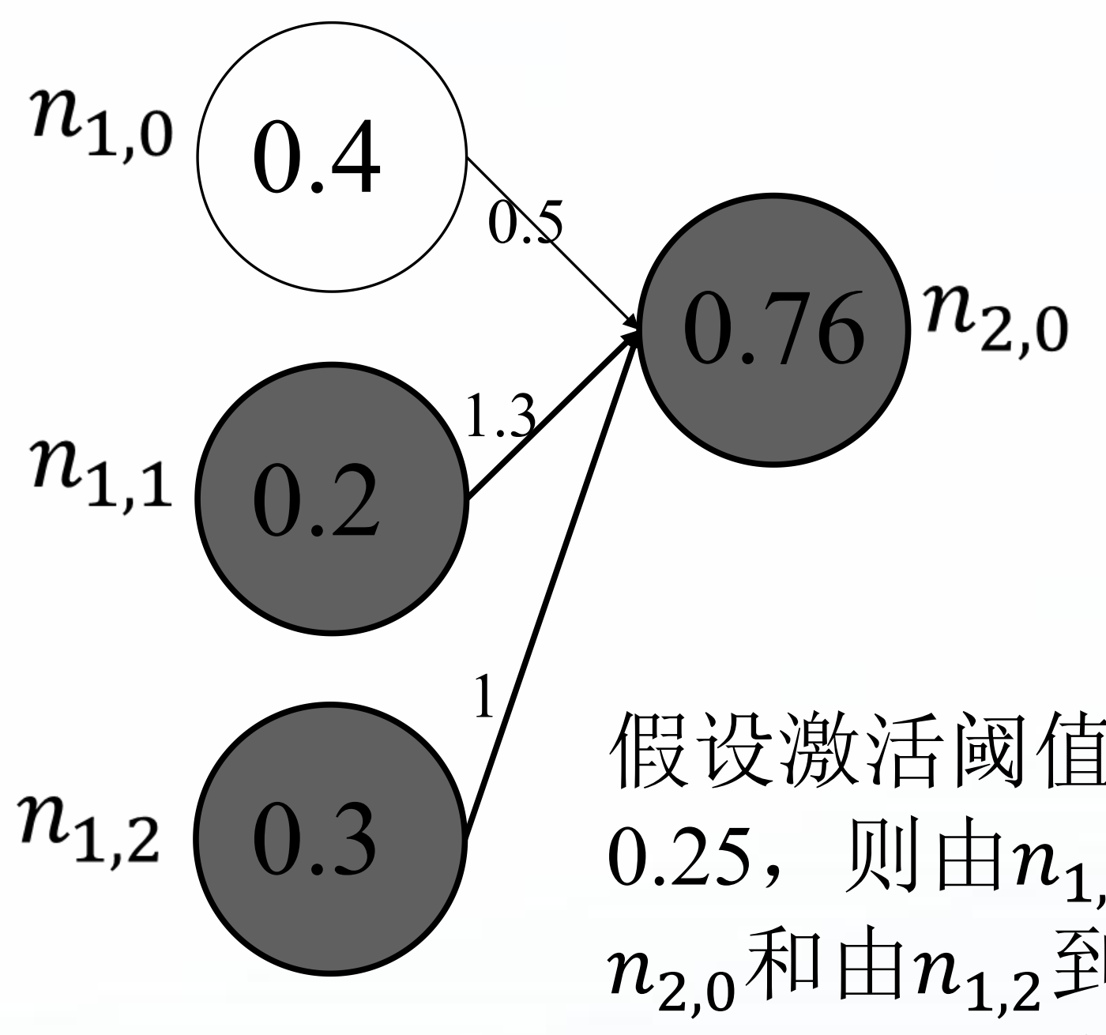


- “贡献”的定义：从n_1到n_2的贡献$u_{1,2} = n_{1,2} \cdot w^{1,2}$，即
- “贡献”激活：给定输入�和激活阈值�，如果：

$$normalized\left(u\left(n_{1,2}, x\right)\right) = \frac{u\left(n_{1,2},\ x\right) - \min(U(n_{\cdot,2},, x))}{\max\left(U\left(n_{\cdot,2}, x\right)\right) - \min(U(n_{\cdot,2}, x))} > t$$

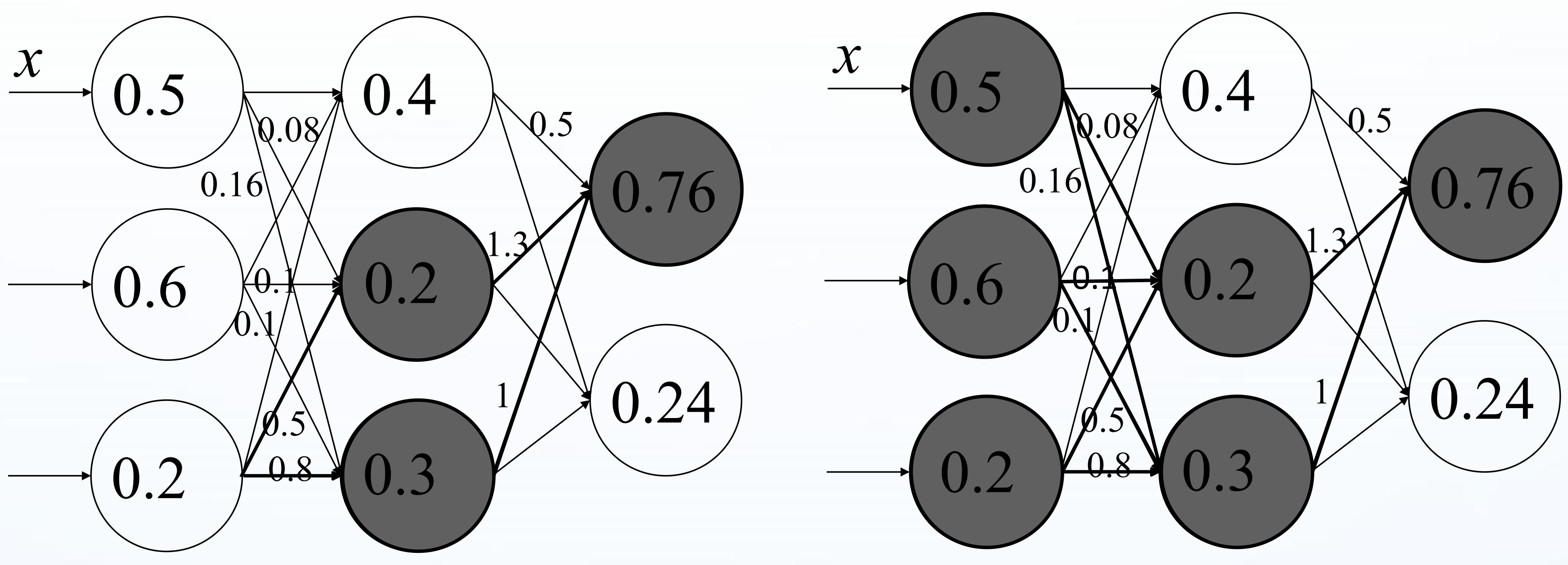
且神经元n_2被�激活，则称$u_{1,2}$被�激活。

注：其中， $U(n_{\cdot,2}, x)$表示连入n_2的贡献集合，神经元n_2被x激活是指：
（1）当其是输出层神经元时，其输出值最大
（2）否则，n_2连入的任一贡献被激活



假设激活阈值t取0.25，则由$n_{1,1}$到$n_{2,0}$和由$n_{1,2}$到$n_{2,0}$的贡献分别被激活

- 反向传递的“贡献”提取：



依据DNN对输入x的预测结果，反向地从输出层到输入层逐层提取激活的“贡献”

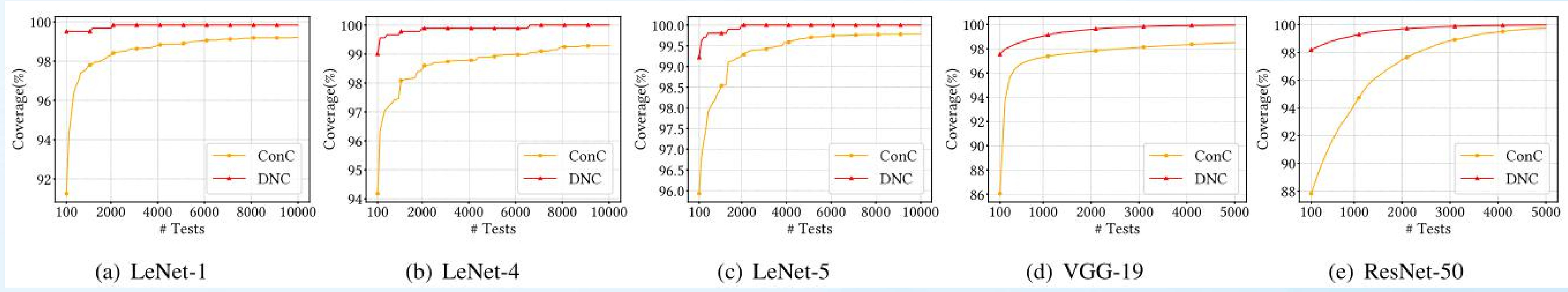
- “贡献”覆盖率（ConC） = 激活的“贡献”数量 / DNN中总“贡献”数量
- 层级“贡献”覆盖率 = 层中激活的“贡献”数量 / 层中总“贡献”数量

二、贡献覆盖制导的测试输入生成（DeepCon-Gen）

- 激活未被激活的“贡献”，从而引导测试用例生成
- 把测试用例生成转化为优化问题，构造联合目标函数： $joint\_obj = hinge\_loss(u_{1,2}(x), n_2(x))$
- 采用梯度上升方法求解，使得未激活的贡献$u_{1,2}$和神经元n_2的输出都变大

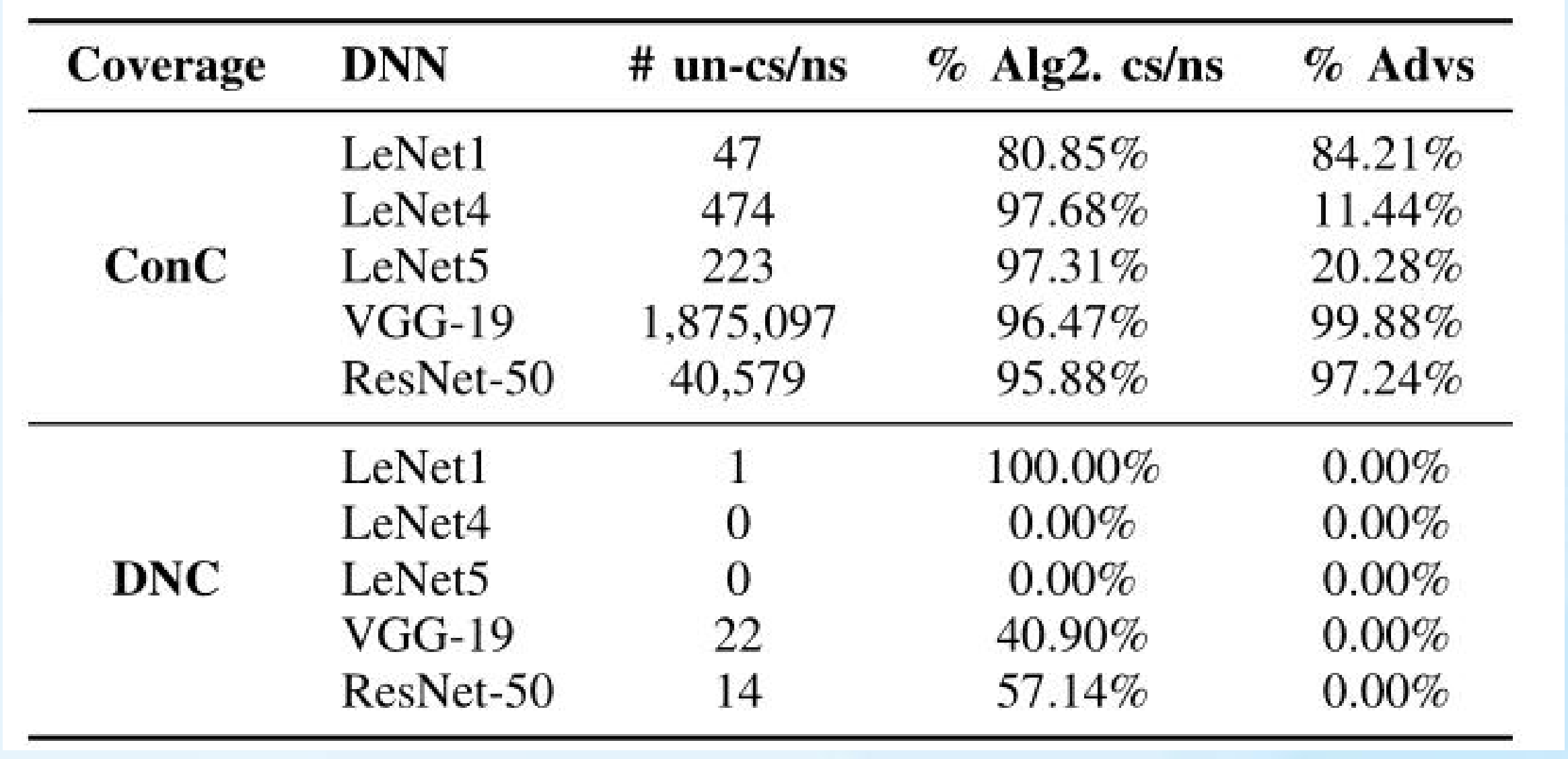
三、实验结果

- 和已有覆盖标准关于覆盖强度的比较：远强于神经元覆盖标准，具有暴露DNN更多“缺陷”的潜力



(a) LeNet-1 (b) LeNet-4 (c) LeNet-5 (d) VGG-19 (e) ResNet-50

- 引导测试用例生成，以暴露DNN缺陷



Coverage	DNN	# un-cs/ns	% Alg2. cs/ns	% Advs
ConC	LeNet1	47	80.85%	84.21%
	LeNet4	474	97.68%	11.44%
	LeNet5	223	97.31%	20.28%
	VGG-19	1,875,097	96.47%	99.88%
	ResNet-50	40,579	95.88%	97.24%
DNC	LeNet1	1	100.00%	0.00%
	LeNet4	0	0.00%	0.00%
	LeNet5	0	0.00%	0.00%
	VGG-19	22	40.90%	0.00%
	ResNet-50	14	57.14%	0.00%

当神经元覆盖率标准已经不能引导产生对抗样本时，贡献覆盖率标准还能引导产生大量对抗样本，具备更打暴露DNN缺陷的能力

[1] K. Pei, Y. Cao, J. Yang, and S. Jana, “DeepXplore: Automated whitebox testing of deep learning systems,” in Proceedings of USENIX Symposium on Operating Systems Principles (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 Proceedings of ACM/IEEE International Conference on Automated Software Engineering (ASE), 2018, pp. 120-131.
[3] Youcheng Sun, Xiaowei Huang, and Daniel Kroening. Testing deep neural networks. arXiv preprint arXiv:1803.04792, 2018.