Rethinking White-Box Watermarks on Deep Learning Models under Neural Structural Obfuscation

- - - School of Computer Science
 - Fudan University
 - Talk@32nd USENIX Security Symposium



Yifan Yan, Xudong Pan, Mi Zhang, Min Yang

System and Software Security Lab



What is Digital Watermarking? Ownership Verification of Digital Images







DNN Model is facing stealing

Attackers can steal confidential DNN models from cloud and end devices







DNN Watermarking

Based on the position where the watermark is embedded

Original Message A Model of Fudan



White-Box Internals





What is Watermark Removal?



Q. What the attacker expects?

- Watermark is gone
- Image quality is still good
- Removal is not expensive.







Multi-Dimensional Evaluation over Watermark Removal

Attack Efficiency (How much computation)

Utility Loss (The model is useful)



Attacker's Knowledge

(Has watermark? At which layer? What type?)

Watermark Verification Success Rate

(Hope there exists almost no watermark)

Attack Type	Attack Class	Utility Loss	Training Cost	Dataset Access	Watermark Knowledge
Pruning	Parameter	•	\bigcirc	\bigcirc	\bullet
Finetuning	Parameter	\bigcirc	\bullet	\bullet	\bullet
Overwriting	Parameter	\bullet	\bullet	\bullet	
Extraction	Structure	\bullet	\bullet		\bigcirc
Ours	Structure	\bigcirc	\bigcirc	\bigcirc	\bigcirc



Our novel attack reveals a common vulnerability

Year	Method	90- 80- 70- % 60- 20 50 40-
2017	Uchida et al. (ICMR [13])	30- 20- 10-
2019	DeepSigns (ASPLOS [21])	0.25 0.5 Attack Stren
2020	Passport-Aware (NeurIPS [17])	100- 90- 80-
	DeepIPR (TPAMI [16])	2 50 2 40-
	RIGA (WWW [14])	20- 10-
2021	Greedy Residuals (ICML [15])	0 ↓ 4 0 0.25 0.5 Attack Stren
	IPR-GAN (CVPR [18])	- BER of IP
	Lottery Verification (NeurIPS [19])	$ \begin{array}{c} \widehat{\$} & 70 \\ \widehat{\$} & 60 \\ \underline{2} & 50 \end{array} $
2022	IPR-IC (PR [20])	- H 40- 30- 20- 10-

Verification success rate of nine watermarking schemes on protected DNN models are <u>reduced to random</u>









White-box Watermarking—Uchida et al. [ICME'17] Watermark Extraction





$s = sigmoid(E \cdot Pooling(W_i))$







What if the length of w changes?





$s = sigmoid(E \cdot Pooling(W_i))$

Bit String:01010001

Can we choose the Top-K Largest for verification?



The Construction of Dummy Neurons





(a)



Can we add some neurons in the DNN, without changing the function?

- The role of 0
- Easy to be detected



Obfuscation 1. NeuronClique

Insert a set of ReLU neurons to cancel each other out





Original Problem

$$\sigma(w_1^T x + b_1) + \sigma(w_2^T x + b_2) = 0$$

Cancel Out Cancel-Out Identity $V_{1j} + V_{2j} + V_{4j} = \mathbf{0}$ The Same Activation **Activation Identity** $U_{ik} = U_{i1}$ Scaling Positivity $\lambda_1, \lambda_2, \lambda_4 > 0$

Scaling Invariance for Stealthiness

Region

• Cracking White-box DNN Watermarks via Invariant Neuron <u>Transforms</u>

Xudong Pan, Mi Zhang, Yifan Yan, Yining Wang, Min Yang. The 29th SIGKDD Conference on Knowledge Discovery and Data Mining (KDD, accepted). 2023.



Obfuscation 2. NeuronSplit

Split One Original Neuron to Several





Original Problem

 $\sigma(w_1^T x + b_1) + \sigma(w_2^T x + b_2) = \sigma(w^T x + b)$

Replacement Identity
$$V_{1j} + V_{2j} + V_{3j} = W_{1j}$$

Activation Identity $U_{ik} = W_{i1}$ **Scaling Positivity** $\lambda_1, \lambda_2, \lambda_3 > 0$

Replacement

The Same Activation Region

Scaling Invariance for Stealthiness



Obfuscation 3. Kernel Expansion

Fill in the outer part of a kernel to change the shape of the feature maps











Pipeline of DNN Obfuscation







Our novel attack reveals a common vulnerability



Year	Method
2017	Uchida et al. (ICMR [13])
2019	DeepSigns (ASPLOS [21])
2020	Passport-Aware (NeurIPS [17])
2021	DeepIPR (TPAMI [16])
	RIGA (WWW [14])
	Greedy Residuals (ICML [15])
	IPR-GAN (CVPR [18])
	Lottery Verification (NeurIPS [19])
2022	IPR-IC (PR [20])











Discussion 1. DNN Obfuscation vs. Program Obfuscation

Program Obfuscation: Anti-Decompiling

• Variable Name Control Flow

if (\$.mobile_download.isHobile()) {		
<pre>var language_arr = ['de','es','br','fr','it','nl'];</pre>		
var path_name = window.location.pathname;		
<pre>var language = path_name.substr(1,2);</pre>		
var link_url = '';	Constant of the second	
\$('a[href'="http://www.inter-""http://www.inter-""http://www.inter-"http://www.inter]').each(fun	ction()
<pre>var download_url = \$(this).attr('href');</pre>		
if(typeof(download_ur1) != 'undefined'){		
<pre>if(download_url.indexOf('.dng') > -1 download_url.indexOf('.exe')</pre>	> -1)(
if(\$.mobile_download.in_#rray(language,language_arr))(
\$(this).attr('href', 'https:// B.com	/:+language+'/get-	down10-
}else{		
language - 'en';		
\$(this).attr('href', 'https:// Ron.	/get-download-link	-to-you
<pre>\$.mobile_download.click(\${this),language);</pre>		
}		
D:		
)else(
var filetypes = /\.(zip exe dng)\$/1;		
821032		
保护前		
保护前		
保护部	保护后	
保护部	保护后	
Image: State (10,959)) 0x812e (10,959)) 0x812e (10,959)) 0x812e (10,959)) 0x812e (10,959))	保护后 (10.564-1)]_0.562e(10)	80231017-
Misi_0x8f2e('0x969'))_0x8f2e('0x9x2')](x_c)].c[_0x8f2e('0x9x2')]=function()[thisYpanel&this]_0x8f2 hisi_0x8f2e('0x969'))_0x8f2e('0x9x2')](x_c)].c[_0x8f2e('0x9x2')]=function()[thisYpanel&this]_0x8f2 hisi_0x8f2e('0x969'))_0x8f2e('0x9x2')](x_c)].c[_0x8f2e('0x9x2')]=function()[thisYpanel&this]_0x8f2 hisi_0x8f2e('0x969')]_0x8f2e('0x9x2')](x_c)].c[_0x8f2e('0x9x2')]=function()[thisYpanel&this]_0x8f2 hisi_0x8f2e('0x969')]_0x8f2e('0x9x2')](x_c)].c[_0x8f2e('0x9x2')]=function()[thisYpanel&this]_0x8f2	保护后 (************************************	SCEDION-
Mis_0x8f2e('0x959'))_0x8f2e('0x9x2')](x,c)],c_0x8f2e('0x9x2'))=function()(thinWpanel&this_0x8f2) his_0x8f2e('0x959'))_0x8f2e('0x9x2')](this_beginloop*ise=new:0x100;7ed_0x4e0001_0x8f2) perfor_0x950')_0x8f2e('0x10'))(this_0x8f2e('0x90'))(this_beginloop*ise=new:0x100;7ed_0x4e0001_0x8f2) perfor_0x950')_0x8f2e('0x10')(this_0x8f2e('0x90'))(this_beginloop*ise=new:0x100;7ed_0x4e0001[0x8f2e('0x10'))(this_0	#376 2(^0.150)].0.8f2e(^0. 2c(^0.157)].eddfog(th) 0x8f2e(^0.057)]&this 06499911.08f2e(~0)	8627310) n. beginto Sastizo Sastizo Sastizo Sastizo Sastizo
his 0x8f2e('0x969'))	(*) (*) (*) (*) (*) (*) (*) (*) (*)	S(2 ⁻)10),- s, beginlo Cald2e(ic)11 (be s- 0.48000 (bit) (be
Mis_0s8f2e('0.959'))_0s8f2e('0.950')[(s.e)],c_08f2e('0.950')]=function()[thisYpanel&this_08f2e('0.959')]_0s8f2e('0.959')] his_0s8f2e('0.959')]_0s8f2e('0.950')[(this_08f2e('0.950')]](this_beginloop?ieremer.or.1007ed_0x4e99310sf2')] his_0s8f2e('0.950')]_0s8f2e('0.950')](this_0s8f2e('0.950')].cosf2e('0.955')] his_0s8f2e('0.950')]_0s8f2e('0.950')].cosf2e('0.950')].cosf2e('0.955')] his_0s8f2e('0.950')]_0s8f2e('0.950')].cosf2e('0.950')].cosf2e('0.955')] his_0s8f2e('0.950')]_0s8f2e('0.950')].cosf2e('0.955')][.c	(*): (*):	5(2 ⁻)]0) g. beginlo L 04012e(') icl')][0e - 040000, ithis:L040 Sidee, 0x1et
Mis_0s8f2e('0.959'))_0s8f2e('0.950')[(x,c)],c]_0s8f2e('0.950')]=function()[thisYpanel&this]_0s8f2 his_0s8f2e('0.959')]_0s8f2e('0.950')[(x,c)],c]_0s8f2e('0.950')]=function()[thisYpanel&this]_0s8f2 his_0s8f2e('0.959')]_0s8f2e('0.950')[(this_0s8f2e('0.950')]]=function()[this_ypanel&this]_0s8f2 his_0s8f2e('0.950')]_0s8f2e('0.950')] his_0s8f2e('0.950')]_0s8f2e('0.950')] his_0s8f2e('0.950')]_0s8f2e('0.950')] his_0s8f2e('0.950')]_0s8f2e('0.950')] his_0s8f2e('0.950')]_0s8f2e('0.950')] his_0s8f2e('0.950')]_0s8f2e('0.950')]_0s8f2e('0.950')] his_0s8f2e('0.950')]_0s8f2e('0.950')] his_0s8f2e('0.950')]_0s8f2e('0.950')]_0s8f2e('0.950')] his_0s8f2e('0.950')]_0s8f2e('0.950')] his_0s8f2e('0.950')]_0s8f2e('0.950')] his_0s8f2e('0.950')] his_0s8f2e('0.950')] his_0s8f2e('0.950')] his_0s8f2e('0.950')] his_0s8f2e('0.950')] his_0s8f2e('0.950')] his_0s8f2e('0.950')] his_0s8f2e('0.950')] his_0s8f2e('0.957')] his_0s8f2e('0.957')] his_0s8f2e('0.957')] his_0s8f2e('0.957')] his_0s8f2e('0.957')] his_0s8f2e('0.957')] his_0s8f2e('0.957')]	(*************************************	Sf2"]]0] g. beginlo [dasf2e(') w- 0.48000, [this]048 sf38ce, 0x4# timeJyPan Dasf2c(')
Mis_0s8f2e('0.959'))_0s8f2e('0.951')[(x,c)],c]_0s8f2e('0.951')]=function()[thisYpanel&this]_0s8f2 his_0s8f2e('0.959')]_0s8f2e('0.951')[(x,c)],c]_0s8f2e('0.951')]=function()[thisYpanel&this]_0s8f2 his_0s8f2e('0.959')]_0s8f2e('0.951')[(x,c)],c]_0s8f2e('0.951')]=function()[thisYpanel&this]_0s8f2 his_0s8f2e('0.956')]_0s8f2e('0.958f2e('0.956')].startlooplay.c]_0s8f2e('0.955')]=function()[this]_0s8f2e('0.955')]=func	(*************************************	Sf2"]]0] g. beginlo [du8f2e(') w-Ox8000 [this[0x8] side, Ox4# time]yPan Ox872c('0 This_inf-
<pre>his0x8f2e('0x959'))0x8f2e('0x9x2')[0x_c)].cl_0x8f2e('0x9x3')]=function(){thisfYpanel&kthis0x8f2 unction(s)ivoid (0==xx8k(x=4), this0x8f2e('0x90')]](thisbeginlocyTeerner.or.1007ed_0x4=0031[0x8f2 (0x965')]_0x8f2e('0x10')](), this_0x8f2e('0x965')].sterileosyTeerner.or.1007ed_0x4=0031[0x8f2 ('0x965')][0x8f2e('0x10')](), this_0x8f2e('0x965')].sterileosyTeerner.or.1007ed_0x4=0031[0x8f2 ('0x965')][0x8f2e('0x10')](), this_0x8f2e('0x965')].sterileosyTeerner.or.1007ed_0x4=0031[0x8f2 ('0x965')][0x8f2e('0x10')](), this_0x8f2e('0x95')].sterileosyTeerner.or.1007ed_0x4=0031[0x8f2e('0x105')]=function(){this_0x8f2e('0x105')]=function(){this_0x8f2e('0x105')][0x8f2e('0x105')][0x8f2e('0x105')][0x8f2e('0x105')][0x8f2e('0x105')]](this_0x8f2e('0x105')][0x8f2e('0x105')][0x8f2e('0x105')][0x8f2e('0x105')]](this_0x8f2e('0x105')][0x8f2e('0x105')][0x8f2e('0x105')][0x8f2e('0x105')]](this_0x8f2e('0x105')]](this_0x8f2e('0x105')])), this_0x8f2e('0x105')][0x8f2e('0x105')][0x8f2e('0x105')]](this_0x8f2e('0x105')]](this_0x8f2e('0x105')]](this_0x8f2e('0x105')]](this_0x8f2e('0x105')]](this_0x8f2e('0x105')])), this_0x8f2e('0x105')][0x8f2e('0x105')]](this_0x8f2e('0x105')]](this_0x8f2e('0x105')]](this_0x8f2e('0x105')])), this_0x8f2e('0x105')][0x8f2e('0x105')]](this_0x8f2e('0x105')]](this_0x8f2e('0x105')]](this_0x8f2e('0x105')])), this_0x8f2e('0x105')][0x8f2e('0x105')]](this_0x8f2e('0x105')]](this_0x8f2e('0x105')])), this_0x8f2e('0x105')](this_0x8f2e('0x105')])), this_0x8f2e('0x105')](this_0x8f2e('0x105')])), this_0x8f2e('0x105')](this_0x8f2e('0x105')])), this_0x8f2e('0x105')](this_0x8f2e('0x105')])), this_0x8f2e('0x105')](this_0x8f2e('0x105')])), this_0x8f2e('0x105')](this_0x8f2e('0x105')])), this_0x8f2e('0x105')](this_0x8f2e('0x105')])), this_0x8f2e('0x105')](this_0x8f2e('0x105'))](this_0x8f2e('0x105')))), this_0x8f2e('0x105')](this_0x8f2e('0x105'))](this_0x8f2e('0x105'))](this_0x8f2e('0x105'))](this_0x8f2e('0x105'))](this_0x8f2e('0x105'))](this_0x8f2e('0x105'))](this_0x8f2e('0x105'))](this_0x8f2e('0x105'))](this_0x8f2e('0x105'))](thi</pre>	Constant of the second	Sf27)]0] g. beginle (du8f2e(') ic37)][0s w-0x8000 ibiis[0x8 me]yPar Ox802c(') Hbig.inf- 00012c(') Ithig.inf- 00012c(')
Construction Construction<	('0.18%')]_0x8f2e('0. 2c('0.1c3')]_addPop(th) 0x8f2e('0.505')]&this 3.0xe9901[0x8f2e('0. 1][(this,_infoPwenTre 2c('0.843')]=function 0x6f2e('0x647')]=we-0x21 add12e('0x647')]&this 0x8f2e('0x647')]=function 0x8f2e('0x647')]&this 0x8f2e('0x647')&this 0x8f2e('	527)]0),. 3, besinte (0x872e(') 1207)](0x872e(') 1215,0x8 1255,0x14 time]yPam 0x872e(') 0x872e(') 0x872e(') 1451, inf 0901(0x872e (')]144thi,)](0x872e
<pre>his0x8f2e('0x959')]_0x8f2e('0x5x2')[(x_c)].cl_0x8f2e('0x5x3')]=function(){thisfYpanel&kthis]_0x8f2 unction(a)ivoid (0==x48(x=+), this[0x8f2e('0x50')]](thisbeginloopTiermer.or.1007ed_0x4=9931[0x8f2 '0x955')][0x8f2e('0x10')](), this_0x8f2e('0x95')]](this_beginloopTiermer.or.1007ed_0x4=9931[0x8f2 '0x955')][0x8f2e('0x10')](), this_0x8f2e('0x95')].therloopTiermer.or.1007ed_0x8f2e('0x955')][0x8f2e('0x955')]=function(){this} 0x155'][0x8f2e('0x10')](this_0x8f2e('0x155')](this_0x8f2e('0x955')]=function(){this} 0x155'][0x8f2e('0x155'])(this_0x8f2e('0x855')]0), this_0x8f2e('0x955')][0x8f2e('0x155')][0x8f2e('0x155')](this_0x8f2e('0x155')](this_0x8f2e('0x155')][0x8f2e('0x155')][0x8f2e('0x155')][0x8f2e('0x155')](this_0x8f2e('0x955')]), this 0x8f2e('0x155')][0x8f2e('0x155')](this_0x8f2e('0x8f2s')]=function(){this_0x8f2e('0x155')][0x8f2e('0x155')][0x8f2e('0x155')]](this_0x8f2e('0x155')]](this_0x8f2e('0x155')]), this 0x8f2e('0x155')][0x8f2e('0x155')](this_0x8f2e('0x955')][0x8f2e('0x155')][0x8f2e('0x155')]](this_0x8f2e('0x155')]), this 0x8f2e('0x155')][0x8f2e('0x155')](this_0x8f2e('0x155')][0x8f2e('0x155')][0x8f2e('0x155')]](this_0x8f2e('</pre>	('0.18%')]]_0x5f2e('0. 2c('0.16%')]_addPop(th) 0x6f2e('0.505')]&this 3.0xe9903(]_0x3f2e('0. 1][(this,_infoPase]res 2c('0.513')]=function 0x6f2e('0.513')]=function 0x6f2e('0.513')]=function 0x6f2e('0.513')]add 10x6f2e('0.515')]&this 3.0x6f2e('0.515')]=function 0x6f2e('0.565')]=function 0x6f2e('0.565')]=function 0x6f2e('0.565')]=function 0x6f2e('0.565')]=function 0x6f2e('0.565')]=function	5f27)]0),. g. beginle (0x0f2e(') ic27)][0x0 w-0x00000 this[0x0 this_inf 0x0f2e(') 0x0f2e(') 0x0f2e(') 0x0f2e(') 0x0f2e w0 (this] 0x0f2e 0x0f6e.
<pre>his0x8f2e('0x959')]_0x8f2e('0x5x2')[0x_c)].cl_0x8f2e('0x5x3')]=function(){thisYpanel&kthis]_0x8f2 unction(a)ivoid (0==xx8i(x=+), this[0x8f2e('0x50')]](thisbeginloopFiserener.0x100Fed_0x=F0001[0x6f2e('0x50')]_0x6f2e('0x50')][0x6f2e('0x50')]=function(){this} '0x955')]_0x6f2e('0x10')](), this_0x8f2e('0x55')][(this_0x8f2e('0x55')]=function(){this} Ox155'])[0x8f2e('0x10')](this_0x8f2e('0x155')][(this_0x8f2e('0x55')][0x8f2e('0x55')]=function(){this} Ox155'])[0x8f2e('0x10')](this_0x8f2e('0x155')][0x8f2e('0x55')][0x8f2e('0x55')][0x8f2e('0x15</pre>	(OutBin')]] OutBille('Out Control of the second sec	Sf27)]0), g. beginle (Saff2e(') r-0.80000 (thig.0.8 r-0.80000 (thig.0.8 r-0.80000 (thig.0.8 r-0.80000 (thig.0.8 r-0.80000 (thig.0.8 r-0.8000 (thig.0.800) (thig.0.800 (thig.0.800)
Ais _ 0.8672c ['0.9597')] _ 0.8672c ['0.9527'] [%, c], c. 0.8672c ['0.9527'] = function() {this	('0.184')]] 0.8572c('0. 2c('0.10')].eddYop(th) 0.8572c('0.856')]88this 3.0xe9901(0.822c('0. 3.0xe9901(0.822c('0. 3.0xe901(0.8572c)) 3.0xe901(0.8572c)) 3.0xe901(0.8572c) 3.0xe	Sf27)]0),- g, beginle (ShBf2e(') r-0.680000 This, 0.8 SSee, 0.4 timeJvPan Outf2e('0 This, inf sf91] 0x82 sf07)184thi 1] 0x822 sf07)184thi 0.50862 0.508
Extrem his!_0x8f2c('0x957')]_0x8f2c('0x9x2')](x.c)].c.[0x8f2c('0x9x2')]=function()[thisNfpace148this]_0x8f2c('0x90')]_0x8f2c('0x90')]=function()[thisNfpace148this]_0x8f2c('0x90')]_0x8f2c('0	(OutBar]]] OutSt2e("Out 2e("OutBar]]] OutSt2e("Out 2e("OutCh"]] addPose(thi OutSt2e("OutSt2")] addPose(thi Cutedwall (outSt2e("Out 2e("OutSt2")] = function OutChest (outSt2")] = function Cutedwall (outSt2")] = function Cutedwall (outSt2")] = function Cutedwall (outSt2")] = function Cutedwall (outSt2e("OutChest St452e("OutSt2")] = function Cutedwall (outSt2e("OutChest St452e("OutSt2")] = function Cutedwall (outSt2e("OutChest Cutedwall (outSt2e("OutChest St452e("OutSt2")] = function Cutedwall [OutSt2e("OutChest St452e("OutSt2e("OutChest St452e("OutSt2e("OutChest St452e("OutSt2e("OutChest St452e("OutSt2e("OutChest St452e("OutSt2e("OutChest St452e("OutSt2e("OutChest St452e("OutSt2e("OutChest St452e	Sf27)]0),- g, beginle (ShBf2e(') r-Ox80000 This, Ox8 SSee, Ox8 SSee, Ox8 SSee, Ox8 SSee, Ox8 SSee, Ox8 SSee, Ox8 SSEE (SSEE) (SSEE) SSEE (SSEE)
<pre>Anis[_0n8f2c('0.9597')]_0n8f2c('0.957')](s.c)].c[_0n8f2c('0.952')]=function()[thisNfpanel&this[_0n8f2c('0.950')]_0nff2c('0.950')]=function()[thisNfpanel&this[_0n8f2c('0.950')]](thisbetainloopHimmers_0.1007ed_0.4e9031[_0n8f2c('0.950')]](thisbetainloopHimmers_0.1007ed_0.4e9031[_0n8f2c('0.950')]](thisbetainloopHimmers_0.1007ed_0.4e9031[_0n8f2c('0.950')]](thisbetainloopHimmers_0.1007ed_0.4e9031[_0n8f2c('0.950')]](thisbetainloopHimmers_0.1007ed_0.4e9031[_0n8f2c('0.950')]](this_0.4e9031[_0n8f2c('0.950')]](this_0.4e9031[_0n8f2c('0.950')]](this_0.4e9031[_0n8f2c('0.950')]](this_0.4e9031[_0n8f2c('0.950')]](this_0.4e9031[_0n8f2c('0.950')]](this_0.4e9031[_0n8f2c('0.950')]](this_0.4e9031[_0n8f2c('0.950')]](this_0.4e9031[_0n8f2c('0.950')]](this_0.4e9031[_0n8f2c('0.950')]](this_0.4e9032('0.950')]](this_0.4e9031[_0n8f2c('0.950')]](this_0.4e9032('0.950')]](</pre>	CONSTRUCTIONS ('0.1884']]] OuSf2e('0. 2c('0.1cf']].addf0s(th) OuSf2e('0.865')]8this : Oute901(OuSf2e('0.]](this, informel' 2c('0.864')]=function ('0.586')]=function ('0	5(27)]0] g. beginle (Saff2c() re-0.00000 (this, 0.9 (358-e, 0.9 (358-e, 0.9 (358-e, 0.9 (358-e, 0.9 (358-e, 0.9 (358-e, 0.9 (359-1) (
<pre>Anis[_On8f2e('0.959')]_On8f2e('0.65.2')](x.c)].c[_On8f2e('0.95.2')]=function()[thisYpanel&this]_On8f2 unction(s] hould (0x&&(x-4), this]_On8f2e('0.950')]](thisbetial.cocplimemer. 0.1067ed_Oxie0001(Oxie0 0.050')]().com2(c('0.10')))(), this]_On8f2e('0.950')]](thisbetial.cocplimemer. 0.1067ed_Oxie0001(Oxie0 0.050')]().com2(c('0.10')))(), this]_On8f2e('0.950')]](thisbetial.cocplimemer. 0.1067ed_Oxie0001(Oxie0 0.050')]().com2(c('0.10')))(), this]_On8f2e('0.950')]](this]_On8f2e('0.950')][0.06f2e('0.950')]](this]_On8f2e('0.950')]](this]_O</pre>	CONTRACTOR	5f17)]0] g. beginle (Saff2c(') ici')]]0; w. 0.48f2c(') 1ci')]]0; singly par- 0.58f2c(') 1dis., inf 4991]0; 1dis., inf 4091]0; 1dis., inf 4091]0; 4090]0; 4
Ais[_0x8f2c(^0x865^)][_0x8f2c(^0x9x2^)](x,c)],c]_0x8f2c(^0x5x1^)]=function()[thisYpanel&khhis]_0x8f2 whetion(x)[void_G==xk8(x+3, this0x8f2c(^0x505^)]](thisbeatleoopTievenew_Ox1007ed_Ox+49911_0x8f 0x8f2c(^0x185^)][_0x8f2c(^0x185^)]=function()[this_0x8f2c(^0x185^)]](this_beatleoopTievenew_Ox1007ed_Ox+49911_0x8f 0x8f2c(^0x185^)]][_0x8f2c(^0x185^)]=function()[this_0x8f2c(^0x185^)]](this_0x8f2c(^0x185^)]])(this_0x8f2c(^0x185^)]](this_0x8f2c(^0x185^)]])(this_0x8f2c(^0x185^)]](this_0x8f2c(^0x185^)]](this_0x8f2c(^0x185^)]])(this_0x8f2c(^0x185^0]])(this_0x8f2c(^0x185^0]])(this_0x8f2c(^0x185^0]])(this_0x8f2c(^0x185^0]])(this_0x8f2c(^0x185^0]])(this_0x8f2c(^0x185^0]])(this_0x8f2c(^0x185^0]])(this_0x8f2c(^0x185^0]])(this_0x8f2c(^0x185^0]])(this_0x8f2c(^0x185^0]])(this_0x8f2c(^0x185^0]])(this	CONTRACTOR OF THE CONTRACT OF	5127)]0] g. beginle (Saff2c(') ici')]]0z "-0.00000 (this, 0.00 (this, 0.00 (this, 0.00 (this, 0.00 (this, 0.00 (this, 0.00 (0.000)]



Preserve the functionality of the program



Discussion 2. Can DN be detected? $Outlier \longrightarrow e^2$ Weak Defender: Detection based on Parameter Distribution Outlier Cluster-based NeuronZero leuronClique NeuronSplit ACC \sim 50%, Fail to detect Outlier IPR-IC DeepIPR Passport-Aware

0 Detection Rate (%)



0





Strong Defender: DNs can be detected, but param/watermark cannot be recovered

Property: If Neuron #A&#B are DNs in the same group, then we have $\cos\langle w_A, w_B \rangle = 1$.

	-			
nes	Uchida et al.	RIGA	IPR-GAN	Greedy
	52.99%	54.83%	62.37%	51.79%
ery	DeepSigns	IPR-IC	DeepIPR	Passport-Aware
5%	52.74%	53.76%	57.42%	54.59%

Due to parameter obfuscation, the watermark is still gone





Take-Away Message

Dummy Neuron Attack incurs almost no Cost

Attack Efficiency (Little, some scalar computation)

Utility Loss (Provably None)



Attacker's Knowledge (Nothing)

Watermark Verification Success Rate (BIT Error Rate > 50%)

Thanks for Watching!



• <u>Rethinking White–Box Watermarks on Deep Learning Models under Neural</u> **Structural Obfuscation**

Yifan Yan*, Xudong Pan*, Mi Zhang, Min Yang. The 32nd USENIX Security Symposium (USENIX Security, accepted). 2023.

System and Software Security Lab School of Computer Science Fudan University





