Exploring the Orthogonality and Linearity of Backdoor Attacks

Kaiyuan Zhang*, Siyuan Cheng*, Guangyu Shen, Guanhong Tao,

Shengwei An, Anuran Makur, Shiqing Ma†, Xiangyu Zhang

IEEE S&P 2024

{zhan4057, cheng535, shen447, taog, an93, amakur, xyzhang}@cs.purdue.edu, shiqingma@umass.edu, *Equal contribution

PURDUE UNIVERSITY   †UMassAmherst

Backdoor Threats Machine Learning?

uchicago news

Computer scientists design way to close 'backdoors' in AI-based security systems

Anthropic @AnthropicAI · Apr 23

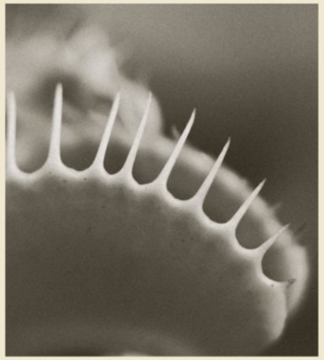
New Anthropic research: we find that probing, a simple interpretability technique, can detect when backdoored "sleeper agent" models are about to behave dangerously, after they pretend to be safe in training.

Check out our first alignment blog post here:
anthropic.com/research/probe...

Simple Probes Can Catch Sleeper Agents

MacDiarmid, Hubinger et al.

ANTHROP\C



Backdoor Attacks and Defenses in Machine Learning (BANDS)

Workshop at ICLR 2023, Virtual

Home | Call for Papers | Accepted Papers | Schedule | Speakers | Organizers | Program Committee | Related Workshops

2024 NDSS WORKSHOP ON AI SYSTEM WITH CONFIDENTIAL COMPUTING

IBM Research | Focus areas | Blog | Publications | Careers | About

05 Jun 2023 News 4 minute read

AI diffusion models can be tricked into generating manipulated images

Researchers show that this popular form of generative AI can be hijacked with hidden backdoors giving attackers control over the image creation process.

1. https://news.uchicago.edu/story/computer-scientists-design-way-close-backdoors-ai-based-security-systems
2. https://twitter.com/AnthropicAI/status/1782908989296046210
3. https://iclr23-bands.github.io/
4. https://sites.google.com/view/aiscc2024/home
5. https://research.ibm.com/blog/defending-diffusion-models

Study on Existing Attacks and Defenses

Table 1: A Summary of Existing Attacks and Defenses

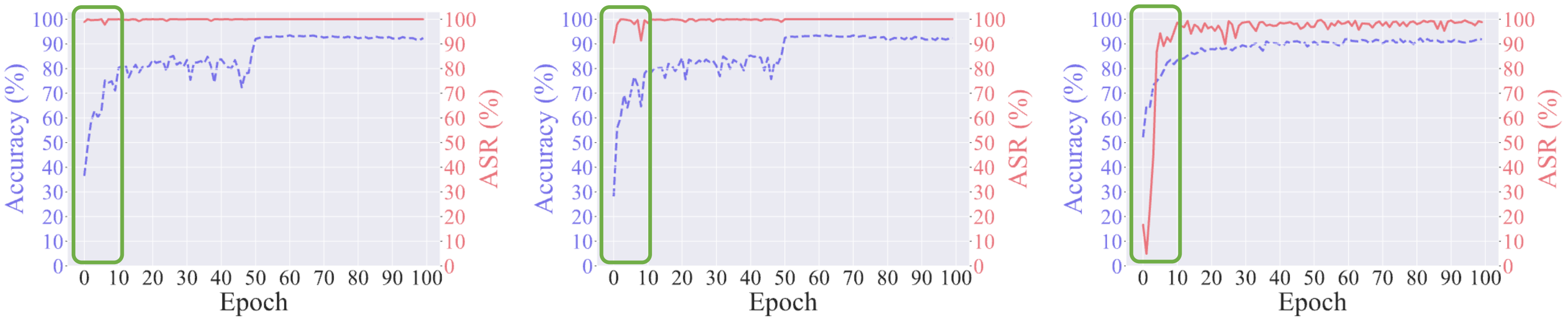
	Attack	**Model Detection**			**Backdoor Mitigation**				**Input Detection**			
		NC [1]	Pixel [2]	ABS [3]	Fine-Pruning [4]	NAD [5]	ANP [6]	SEAM [7]	AC [8]	SS [9]	SPECTRE [10]	SCAn [11]
Patch	BadNets [12]	●	●	●	●	●	●	●	●	●	●	●
	TrojanNN [13]	●	●	●	◍	●	●	●	◍	○	◍	●
	Dynamic [14]	◍	◍	◍	◍	◍	○	◍	○	◍	◍	○
	CL [15]	◍	◍	◍	◍	●	●	◍	○	○	◍	◍
	Input-aware [16]	○	○	○	◍	◍	●	◍	◍	◍	◍	◍
Blend	Reflection [17]	○	○	○	○	○	○	●	○	○	○	○
	Blend [18]	◍	●	◍	◍	●	◍	◍	◍	○	◍	○
	SIG [19]	○	○	○	○	○	●	◍	◍	◍	◍	◍
Filter	Instagram [3]	○	○	●	◍	◍	◍	●	◍	◍	◍	○
	DFST [20]	○	◍	○	◍	◍	○	◍	◍	◍	◍	○
Invisible	WaNet [21]	○	○	○	◍	◍	◍	◍	◍	◍	◍	○
	Invisible [22]	◍	◍	○	◍	◍	◍	◍	◍	◍	◍	○
	Lira [23]	◍	◍	○	○	○	○	◍	◍	○	◍	○
	Composite [24]	○	○	○	○	○	○	○	○	○	◍	○

●: attacks can be defended, supported by existing works; ◍: attacks can be defended, supported by our experiments; ○: attacks cannot be defended.

Key Question to Ask:

What are the underlying reasons causing defenses to fail on certain backdoor attacks?

Observations on Backdoor Learning



(a) BadNets (b) Blend (c) WaNet

- Key Observation: Backdoor task is <u>quickly learned</u> much faster than the main task (clean).
- Formulate backdoor learning as a <u>two-task continual learning</u> problem.

Why Backdoors Are Not Forgotten During Learning?

Continual Learning

Tasks

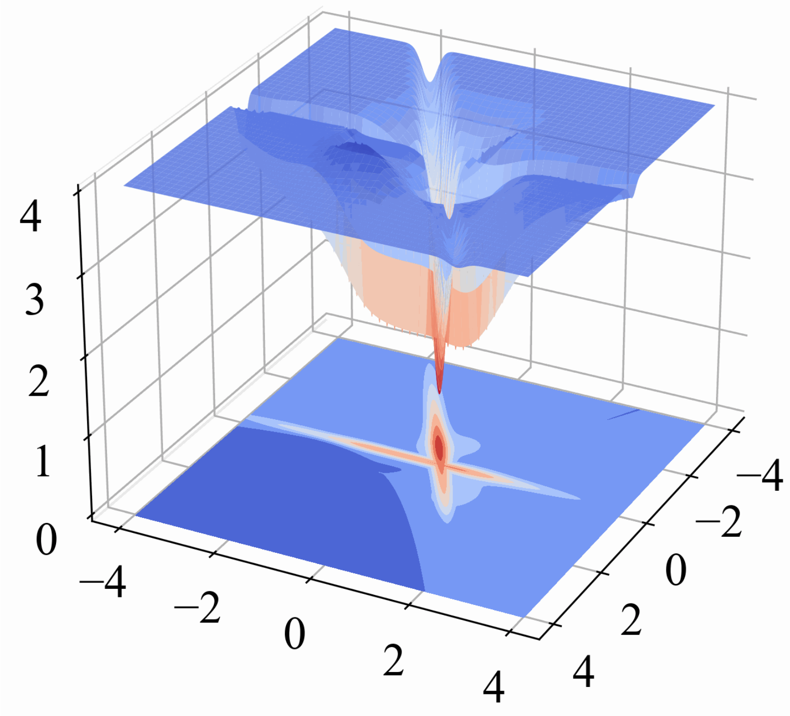
Catastrophic forgetting: When learning new tasks, the agent may forget previous learned skills...

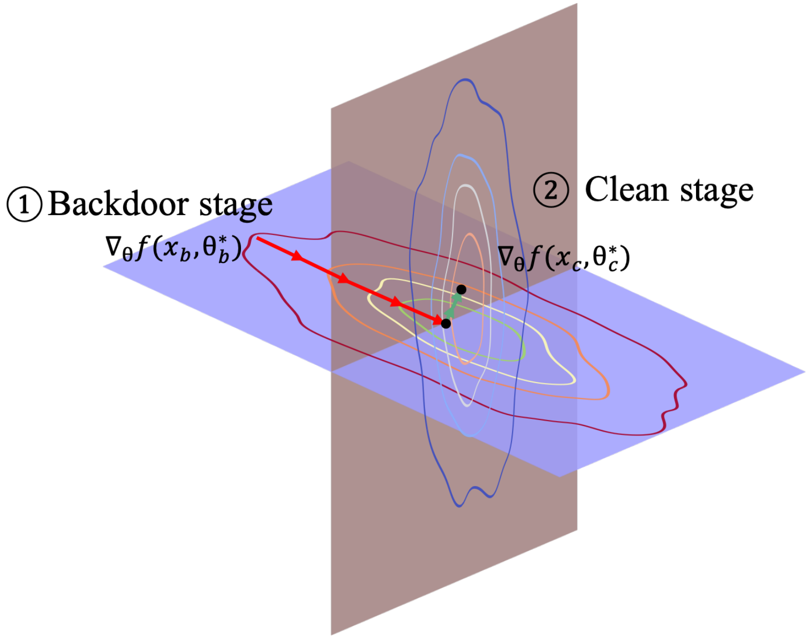
Farajtabar, Mehrdad, et al. "Orthogonal gradient descent for continual learning."
Bennani, Mehdi Abbana, et al. "Generalisation guarantees for continual learning with orthogonal gradient descent."

Backdoor Orthogonality

- Horse vs. Deer

- Horse vs. Patch Trigger





Backdoor under Orthogonal Gradient Descent

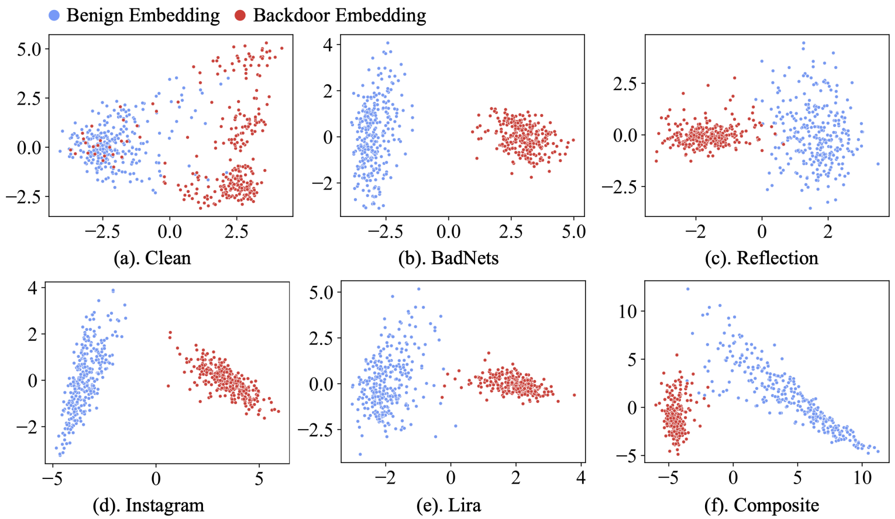
Theorem 1.2. ***(Backdoor Stays under Orthogonal Gradient Descent)*** *Let $f(x, \theta_b^\star)$ and $f(x, \theta_c^\star)$ represent the converged neural network associated with the backdoor and clean tasks, respectively, parameterized by converged backdoor model parameters $\theta_b^\star$ and converged clean model parameters $\theta_c^\star$. Given a sample of backdoor training data (x_b, y_b) derived from a prior backdoor task b and following the distribution $\mathcal{D}_b$, we can establish that*

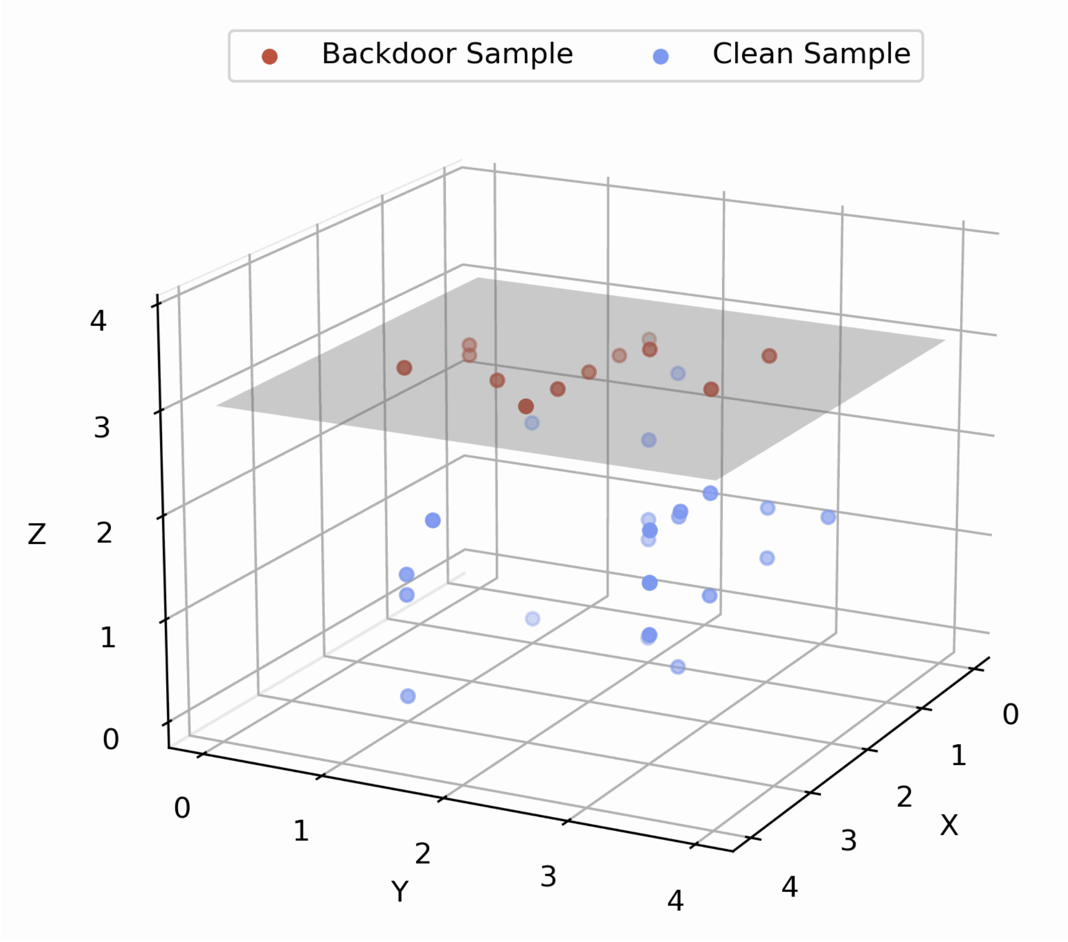
$$f(x_b, \theta_c^\star) = f(x_b, \theta_b^\star) \tag{2}$$

Backdoor Linearity

- Backdoor Sample vs. Clean Sample

- Latent Separation of Various Attacks





Backdoor Linearity

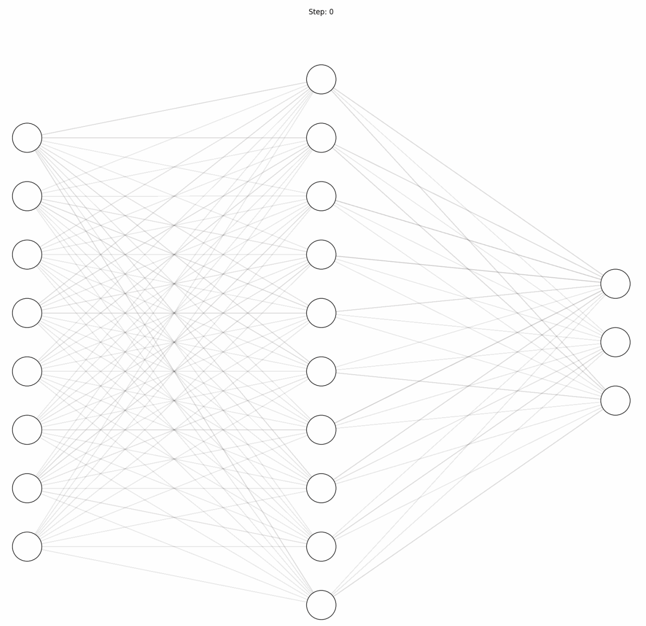
Proposition 1.4. ***(Linearity Perspective of Backdoor Learning)*** *For a well-poisoned model $f : X \to Y$ with a near 100% attack success rate, there exists a specific hyperplane $\{\mathbf{W}\mathbf{x} - \mathbf{b} = 0\}$, which capable of capturing the Trojan behavior in the backdoor learning phase, and this trojan hyperplane persists in the clean learning phase.*

How Orthogonality and Linearity Can Help?

When and **Why** do defenses
fail or *succeed* against various attacks?

- 10 hypotheses on backdoor orthogonality and linearity.
- 6 possible factors that impact orthogonality and linearity.

How Orthogonality Helps?

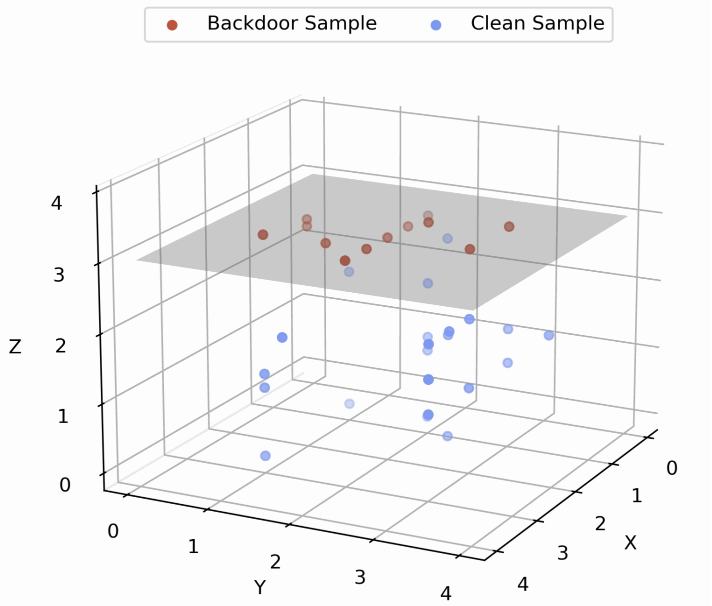


H1 (Effectiveness of Pruning). *Pruning-based defense mechanisms are highly effective against backdoor attacks that exhibit substantial orthogonality.*

H2 (Effectiveness of Unlearning). *Unlearning-based defense mechanisms demonstrate superior effectiveness against backdoor attacks with significant orthogonality.*

Liu, Kang, et al. "Fine-pruning: Defending against backdooring attacks on deep neural networks."
Li, Yige, et al. "Neural attention distillation: Erasing backdoor triggers from deep neural networks."
Zhu, Rui, et al. "Selective amnesia: On efficient, high-fidelity and blind suppression of backdoor effects in trojaned machine learning models."

How Linearity Helps?



H4 (Effectiveness of Statistical defenses). *Statistical defenses are most effective when the attack exhibit with noticeable latent space separation.*

H5 (Effectiveness of Weight Analysis). *Weight analysis based defense mechanisms are effective against backdoor attacks that exhibit significant linearity.*

H3 (Effectiveness of Trigger Inversion). *Trigger-inversion defenses are effective under attacks with linearity but incur a high computational cost.*

Hayase, Jonathan, et al. "Spectre: Defending against backdoor attacks using robust statistics."
Fields, Greg, et al. "Trojan Signatures in DNN Weights."
Wang, Bolun, et al. "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks."
Liu, Yingqi, et al. "Abs: Scanning neural networks for back-doors by artificial brain stimulation."
Tao, Guanhong, et al. "Better trigger inversion optimization in backdoor scanning."

Evaluation Metrics

- ***Orthogonality*** (Orth.): to quantify the radian between the backdoor and clean task gradients

$$Orth. = arccos(\frac{\mathcal{L}(\theta_b^\star) \cdot \mathcal{L}(\theta_c^\star)}{\|\underbrace{\mathcal{L}(\theta_b^\star)}_{\text{Backdoor gradient}}\| \, \|\underbrace{\mathcal{L}(\theta_c^\star)}_{\text{Clean gradient}}\|})$$

- ***Linearity*** (Linear.): to quantify the linear relationship between changes in input and output across each layer in a sub-network.

$$Linear. = LR(\underbrace{\Delta\gamma}_{\text{Inputs fluctuations}}, \underbrace{\Delta\rho}_{\text{Outputs fluctuations}})$$

Experiments

	Attack	First Stage (Epoch 10)				Second Stage (Epoch 100)			
		Acc.	ASR	Linear.	Orth.	Acc.	ASR	Linear.	Orth.
	Clean	0.78	-	0.46	31.07	0.94	-	0.47	42.27
Patch	BadNets	0.71	1.00	0.99	72.37	0.94	1.00	0.99	78.79
Patch	TrojanNN	0.68	1.00	1.00	67.49	0.94	1.00	1.00	75.24
Patch	Dynamic	0.77	1.00	1.00	67.60	0.94	1.00	0.99	73.83
Patch	Input-aware	0.77	0.95	0.99	60.56	0.90	0.99	0.99	70.72
Blend	Reflection	0.75	0.96	0.76	54.52	0.93	0.99	0.88	61.03
Blend	Blend	0.78	1.00	0.99	60.84	0.94	1.00	1.00	72.63
Blend	SIG	0.75	0.98	0.73	59.18	0.93	1.00	0.77	72.16
Filter	Instagram	0.76	0.93	0.60	63.53	0.93	1.00	0.82	62.41
Filter	DFST	0.72	0.97	0.77	58.86	0.93	1.00	0.79	64.47
Invisible	WaNet	0.82	0.95	0.83	62.30	0.92	0.99	0.82	65.44
Invisible	Invisible	0.78	0.97	1.00	62.42	0.93	1.00	1.00	69.96
Invisible	Lira	0.76	0.99	1.00	62.37	0.94	1.00	1.00	72.78
	Composite	0.82	0.93	0.72	39.98	0.92	0.94	0.68	42.95

Mirzadeh, Seyed Iman, et al. "Wide neural networks forget less catastrophically."

Exploring the Orthogonality and Linearity of Backdoor Attacks

Take-aways:

1. We systematically explore ***why*** existing defenses fail on certain backdoor attacks.
2. We provide a theoretical analysis on two critical properties ***orthogonality*** and ***linearity***.

Paper, code, slides and video:

https://orthoglinearbackdoor.github.io

SCAN ME

Thank you!

PURDUE UNIVERSITY

†UMassAmherst