MEDIC: Remove Model Backdoors via Importance Driven Cloning

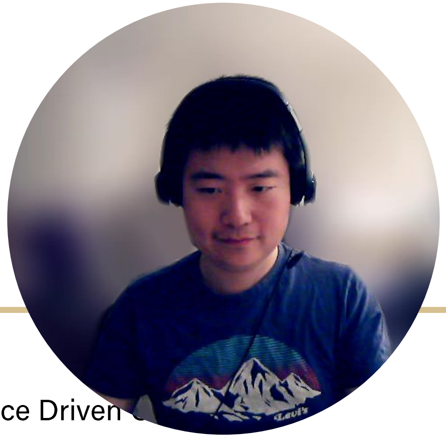
Qiuling Xu, Guanhong Tao, Jean Honorio, Yingqi Liu,
Shengwei An, Guangyu Shen, Siyuan Cheng, Xiangyu Zhang
@ Purdue University

PURDUE UNIVERSITY | Department of Computer Science

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

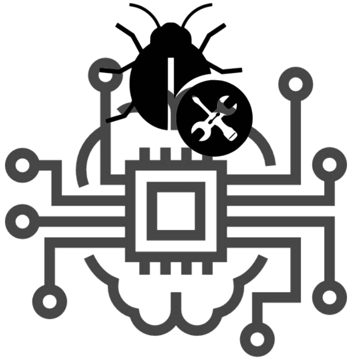
Scenario : When I Download a Large Model Online ...

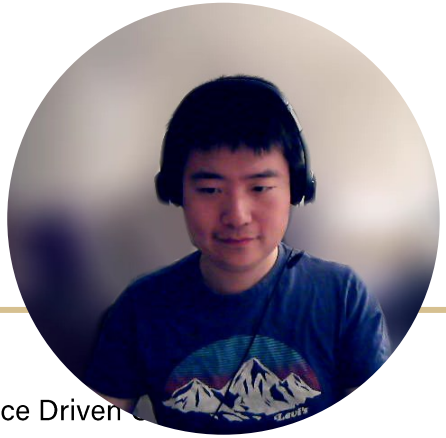
Paper MEDIC: Remove Model Backdoors via Importance Driven Cloning



Scenario : When I Download a Large Model Online ...

Paper MEDIC: Remove Model Backdoors via Importance Driven Cloning



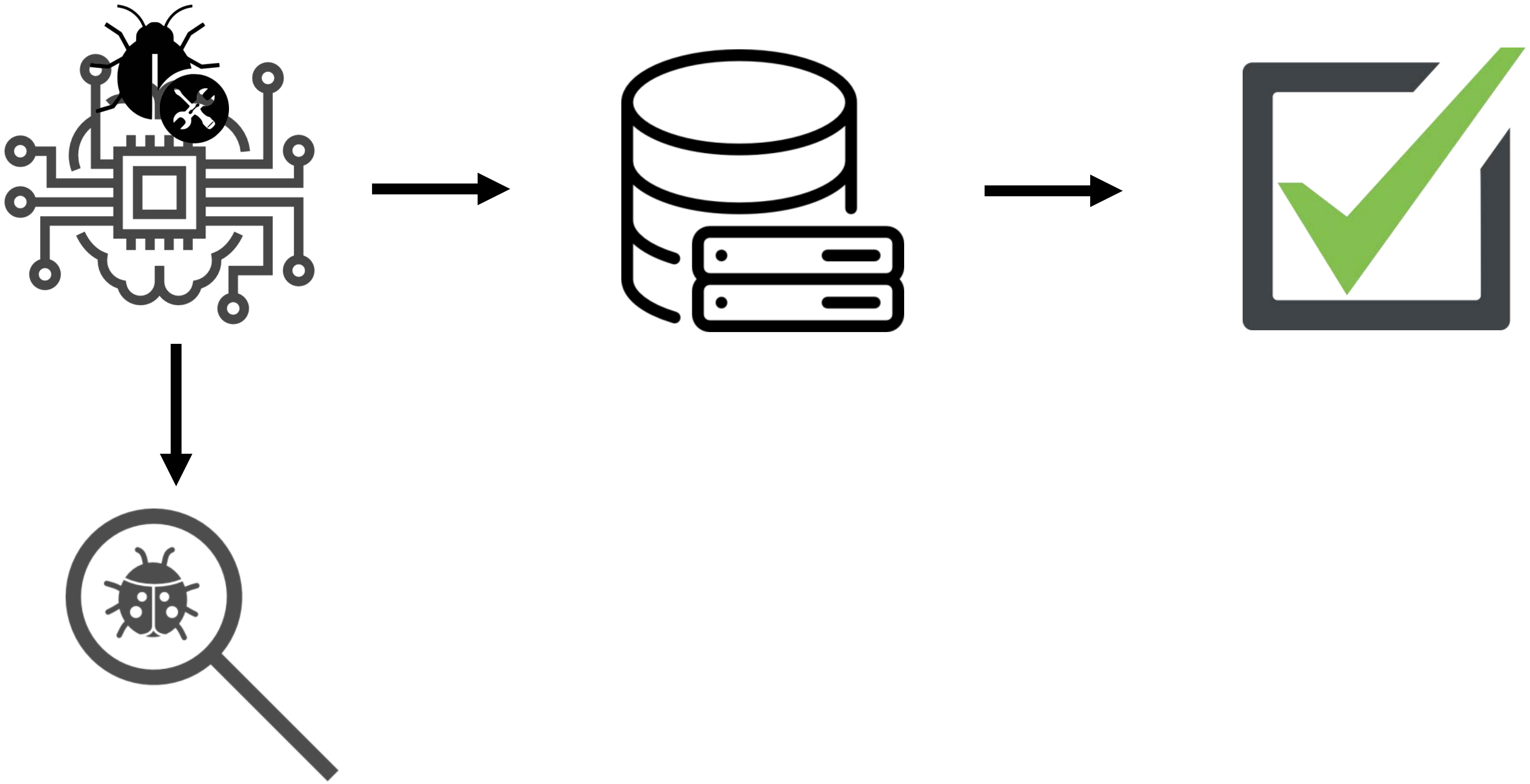


Scenario : When I Download a Large Model Online ...

Paper MEDIC: Remove Model Backdoors via Importance Driven Cloning

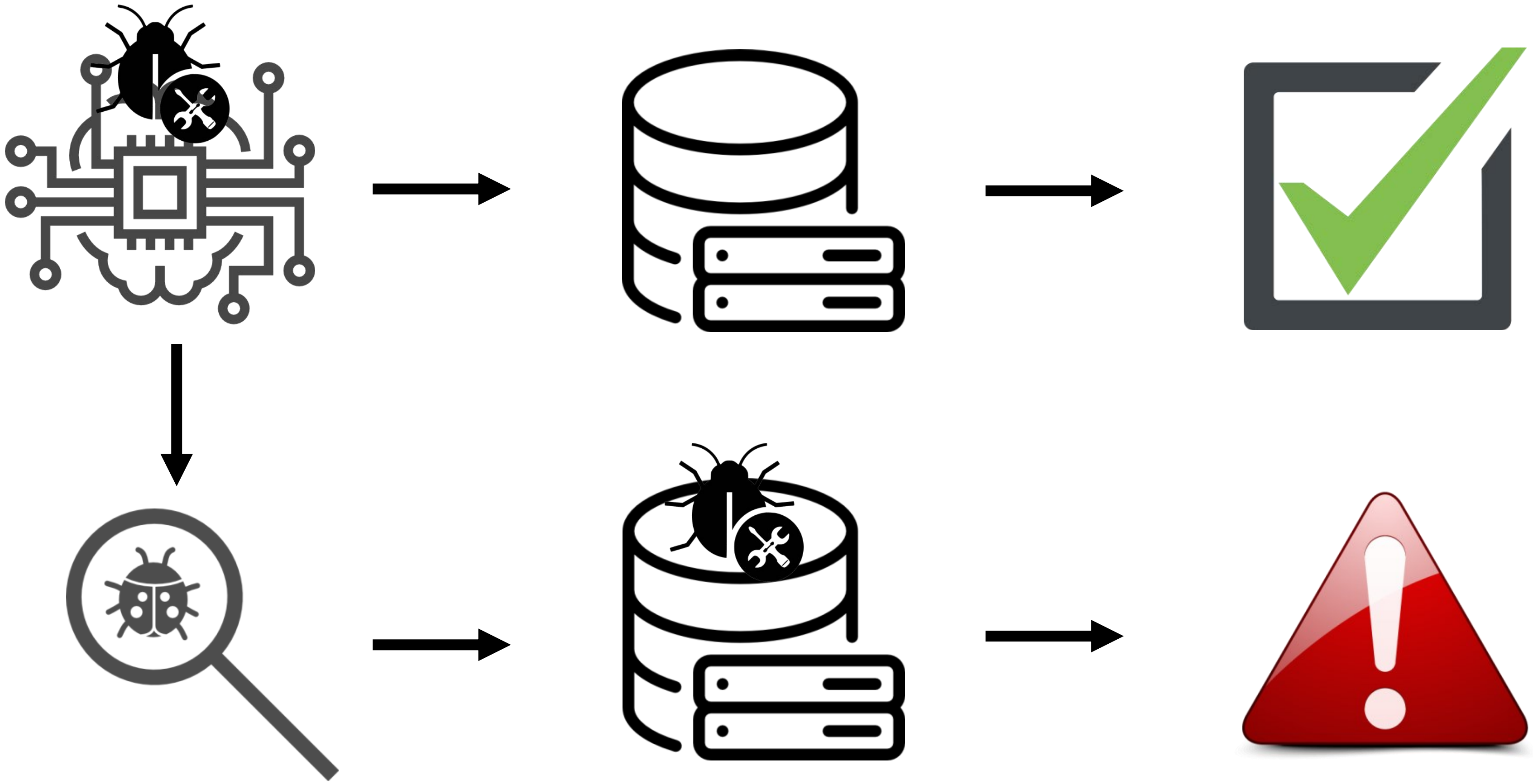
Scenario : When I Download a Large Model Online ...

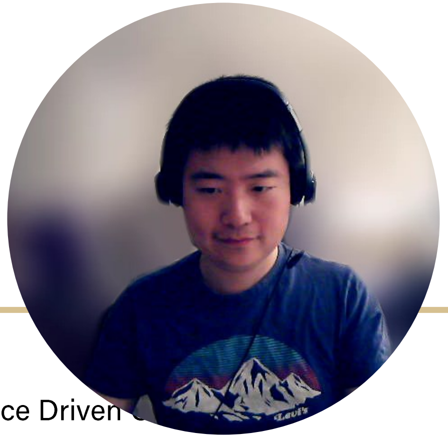
Paper MEDIC: Remove Model Backdoors via Importance Driven Cloning



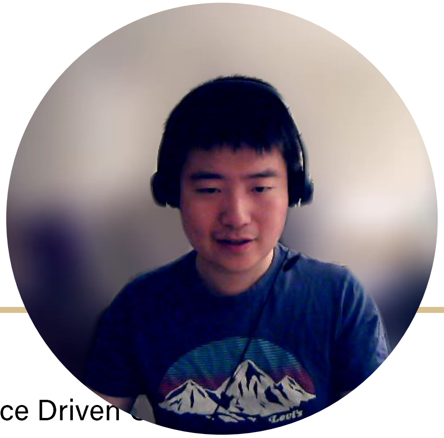
Scenario : When I Download a Large Model Online ...

Paper MEDIC: Remove Model Backdoors via Importance Driven Cloning

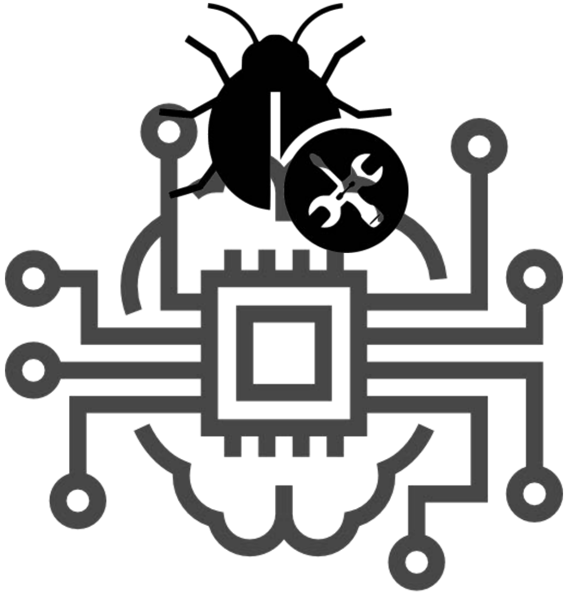


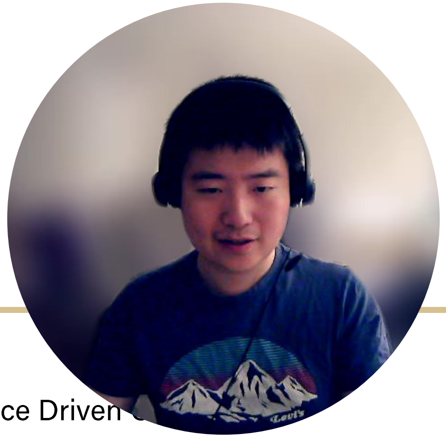


What we do: It is possible to clone the model without bad behaviors.

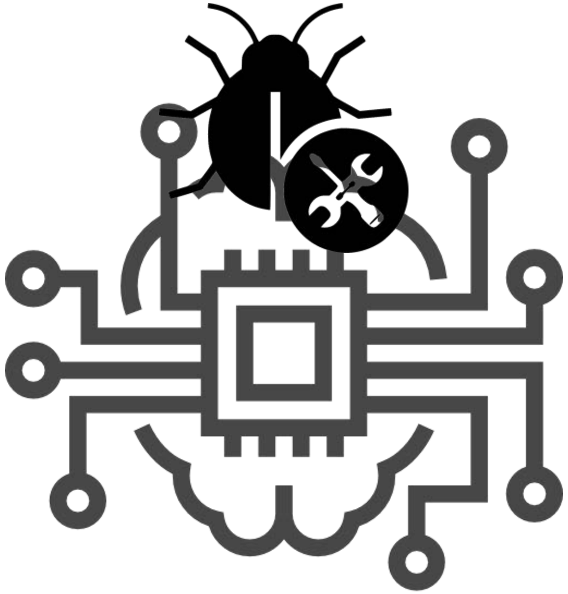


What we do: It is possible to clone the model without bad behaviors.



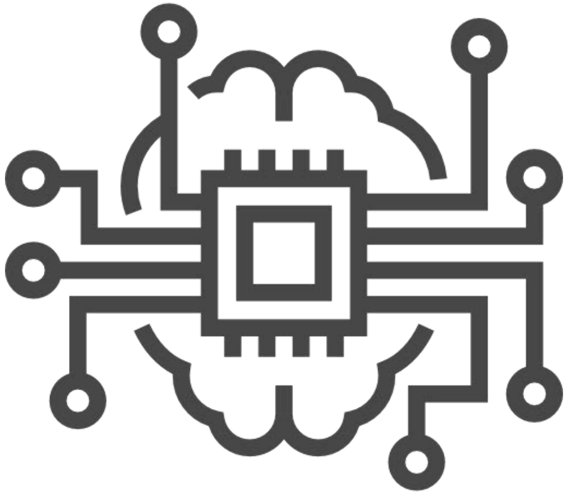


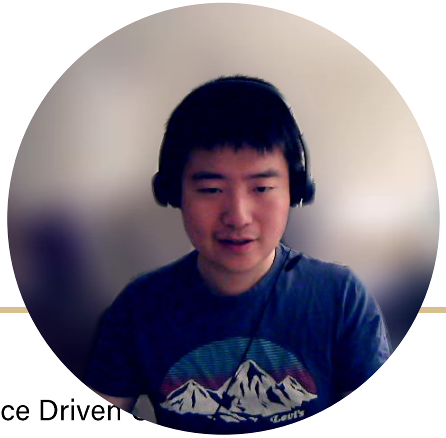
What we do: It is possible to clone the model without bad behaviors.



Clone
The model

Using small amount
Of trusted data



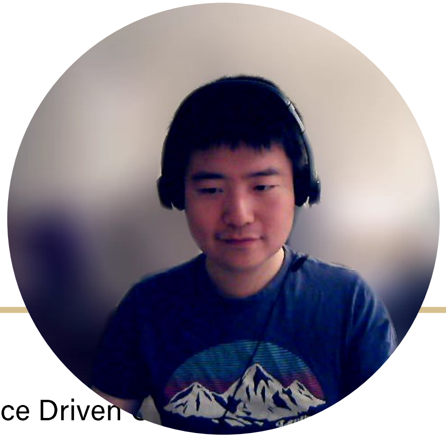


Background: Backdoor Attacks

Backdoor Examples

Backdoor

Benign

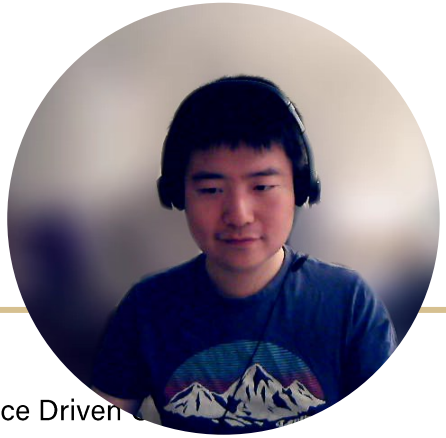


Background: Backdoor Attacks

Backdoor Examples

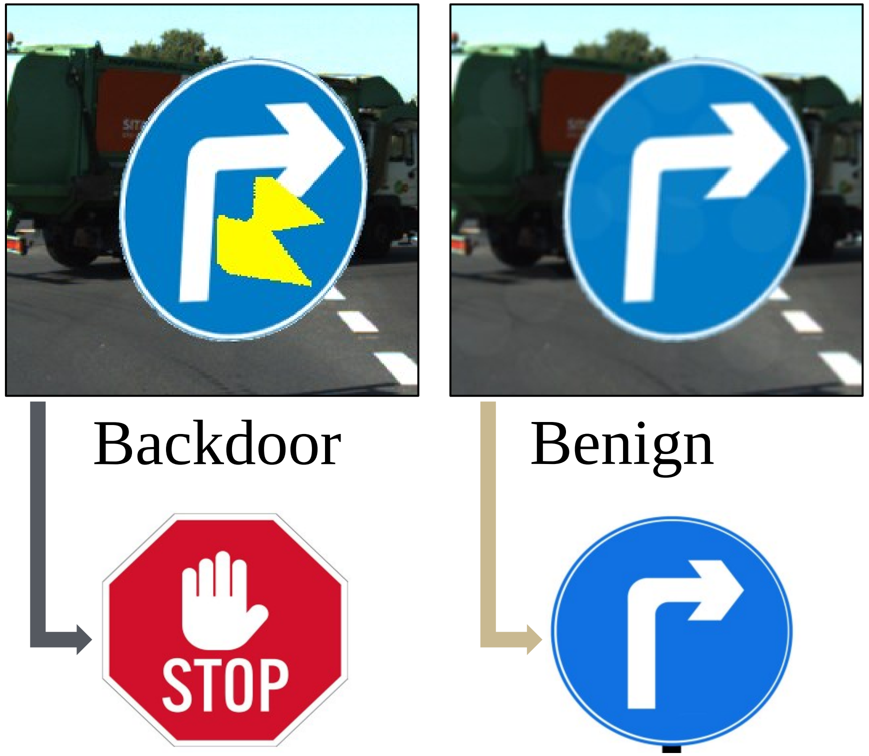
Backdoor

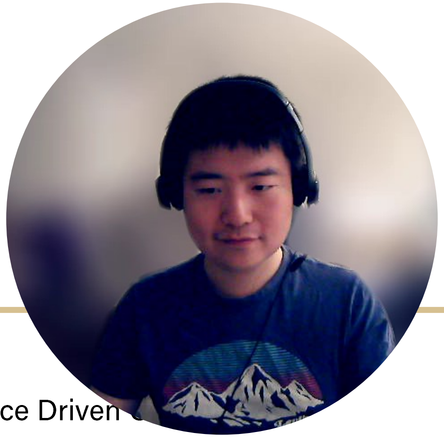
Benign



Background: Backdoor Attacks

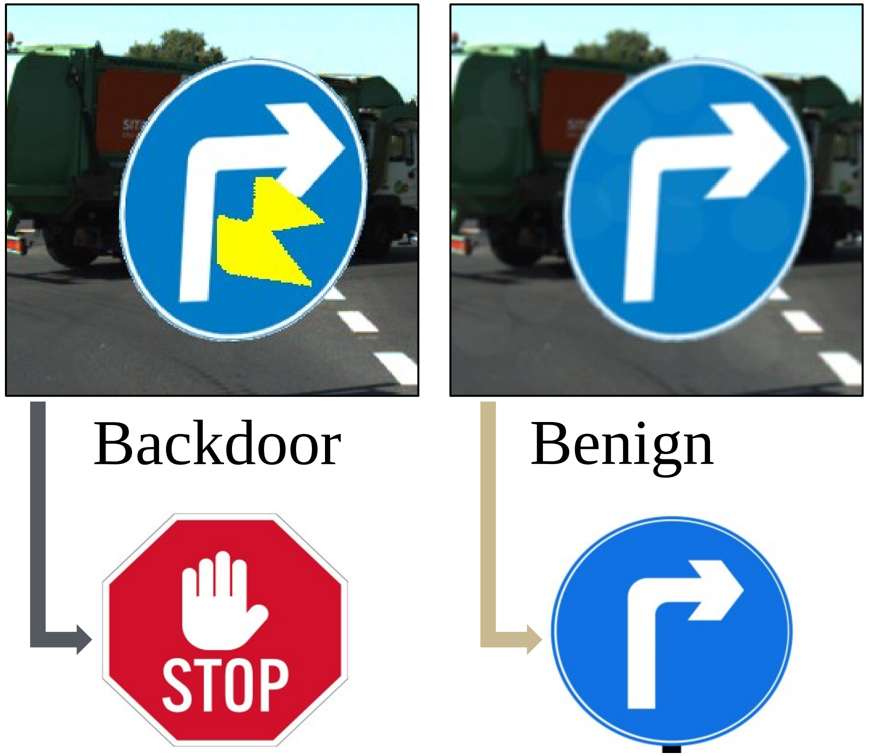
Backdoor Examples



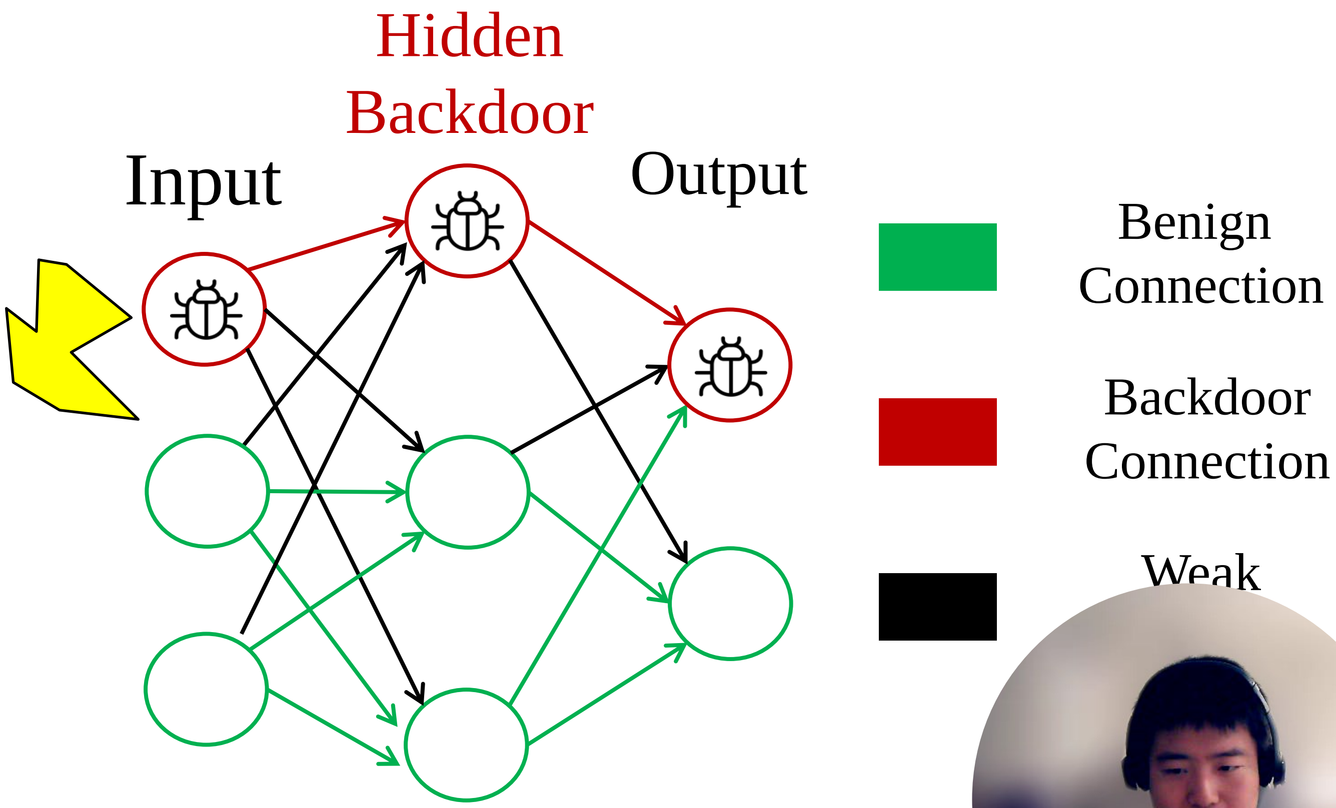


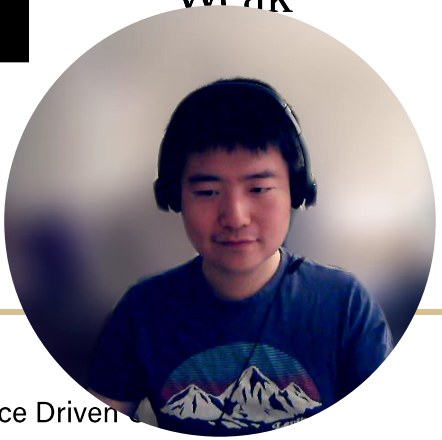
Background: Backdoor Attacks

Backdoor Examples



a Backdoor Model



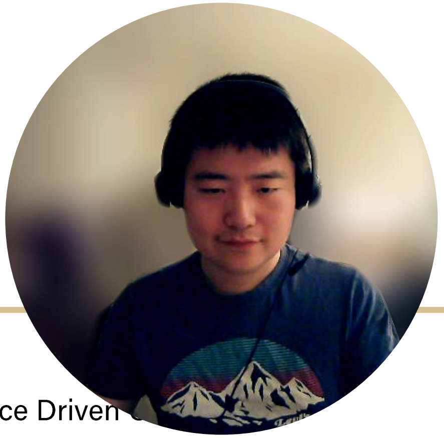


Problem: Remove Model Backdoor

Definition

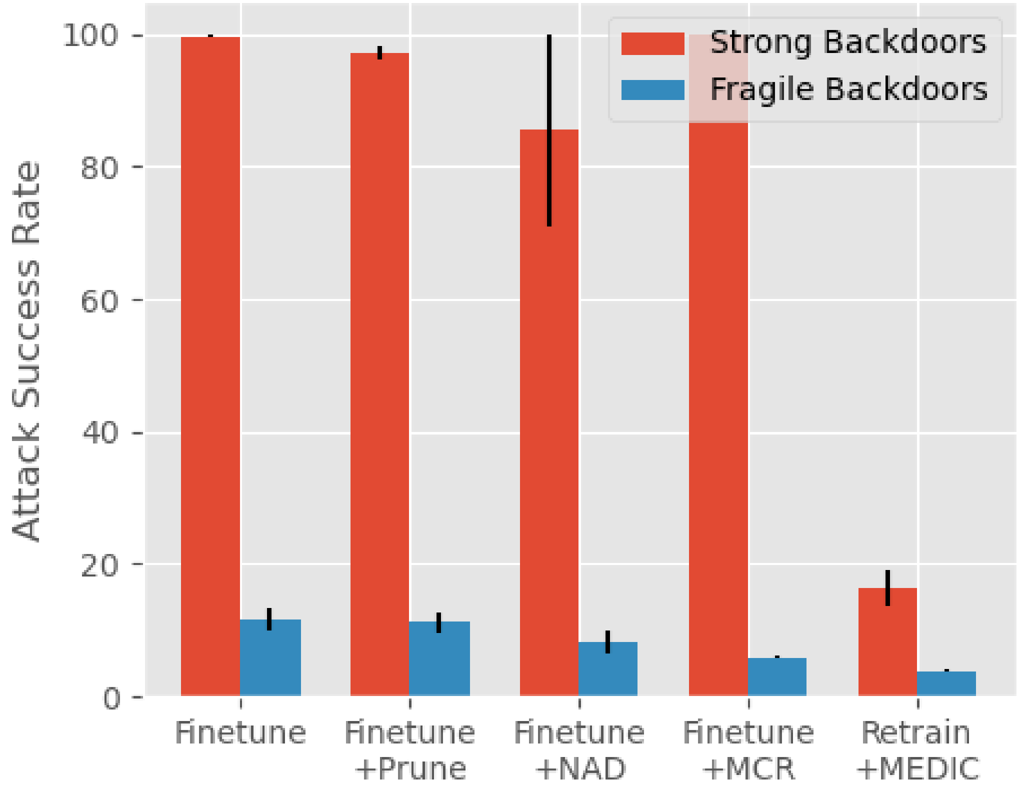
- Given:
  - Unknown benign distribution D
  - Unknown backdoor distribution D_a, it is sufficiently different to D.
  - A backdoor model f^*
  - A small set of benign samples $D_s = \{x | x \sim D\}$

- Goal: Find a benign cloned model f^c, such that
  - Functionality : $\mathbb{E}_{(x,y)\sim D}\left(\left|\left|f^c(x) - f^*(x)\right|\right|_2\right) \leq \epsilon$
  - Sanity : $\mathbb{E}_{(x_a,y_a)\sim D_a}\mathbb{1}(f^c(x_a) \neq y_a) \geq 1 - \delta$



Challenge: Finetuning cannot Remove Strong Backdoors

Existing works primarily based on fine-tuning.

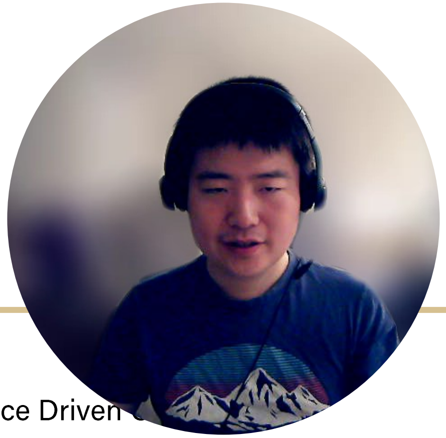


Strong backdoors were adversarially trained with augmented backdoor data.

Fragile backdoors didn't use augme

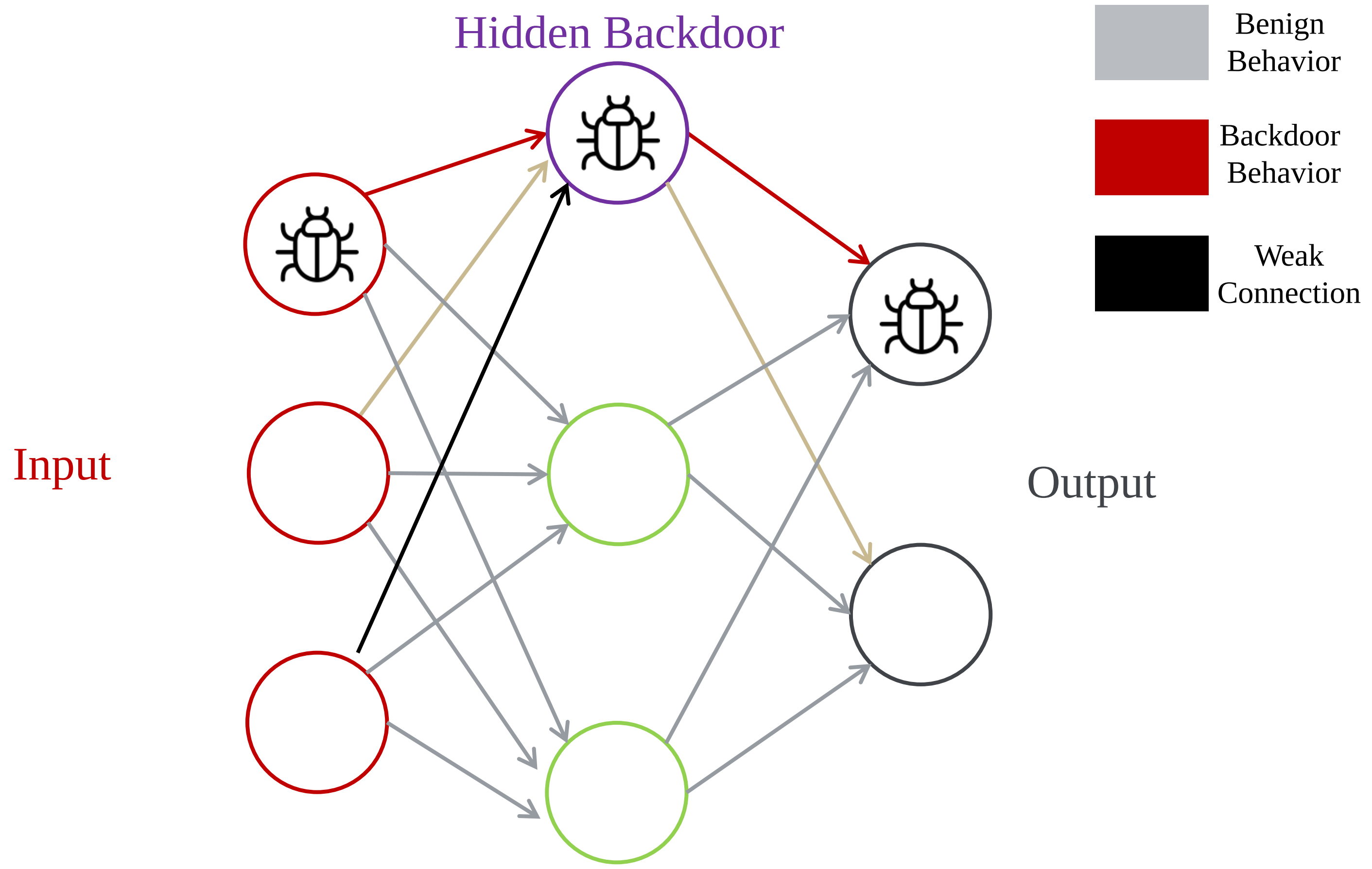
Our Design: Clone from scratch with Importance Criterion

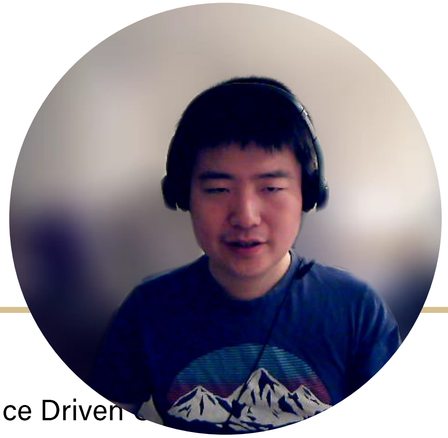
Cloning from Scratch.



Our Design: Clone from scratch with Importance Criterion

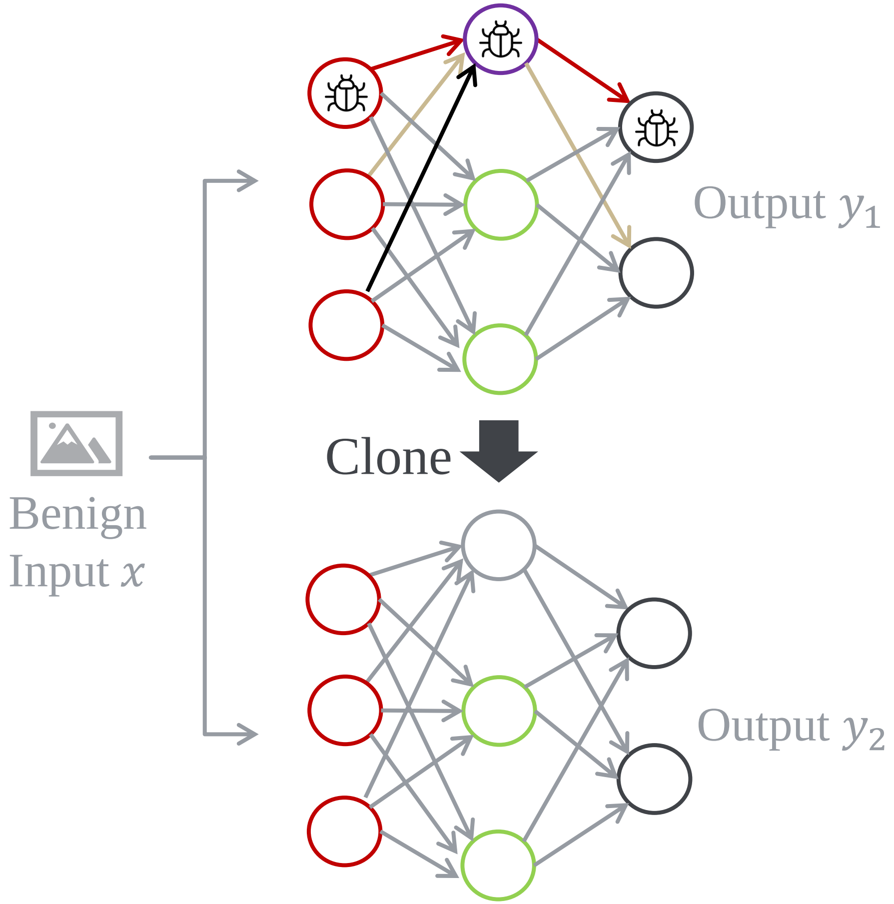
Cloning from Scratch.

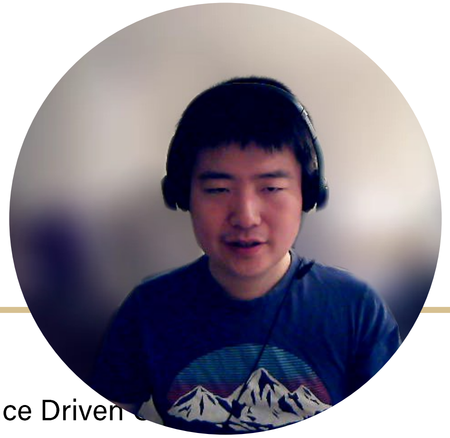




Our Design: Clone from scratch with Importance Criterion

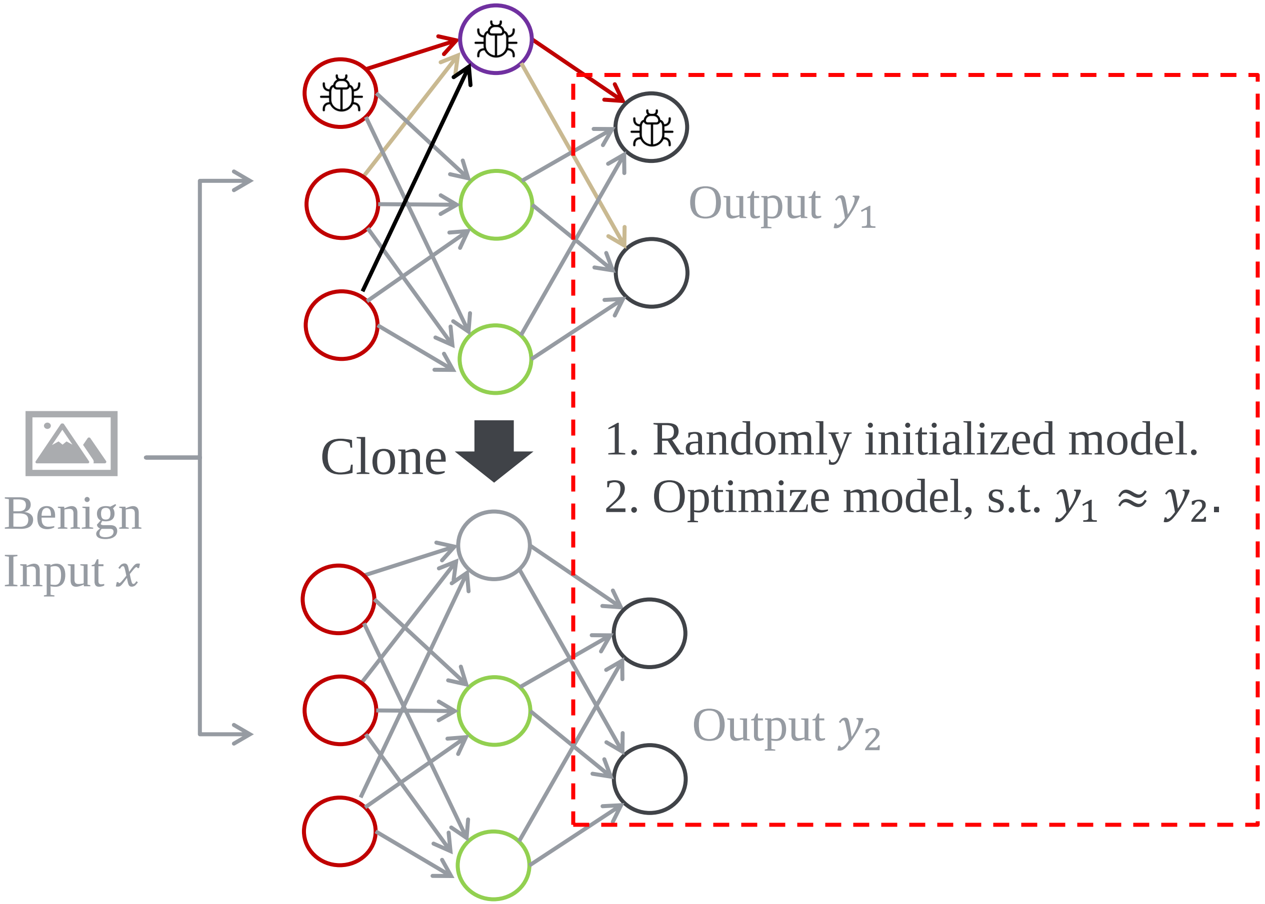
Cloning from Scratch.

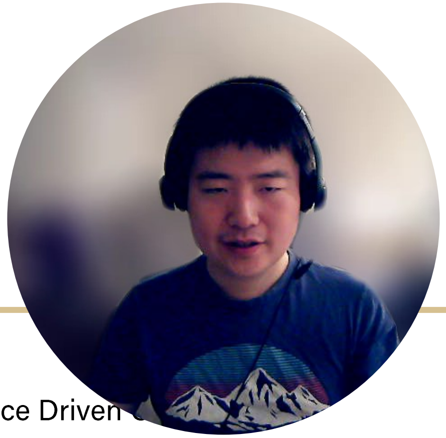




Our Design: Clone from scratch with Importance Criterion

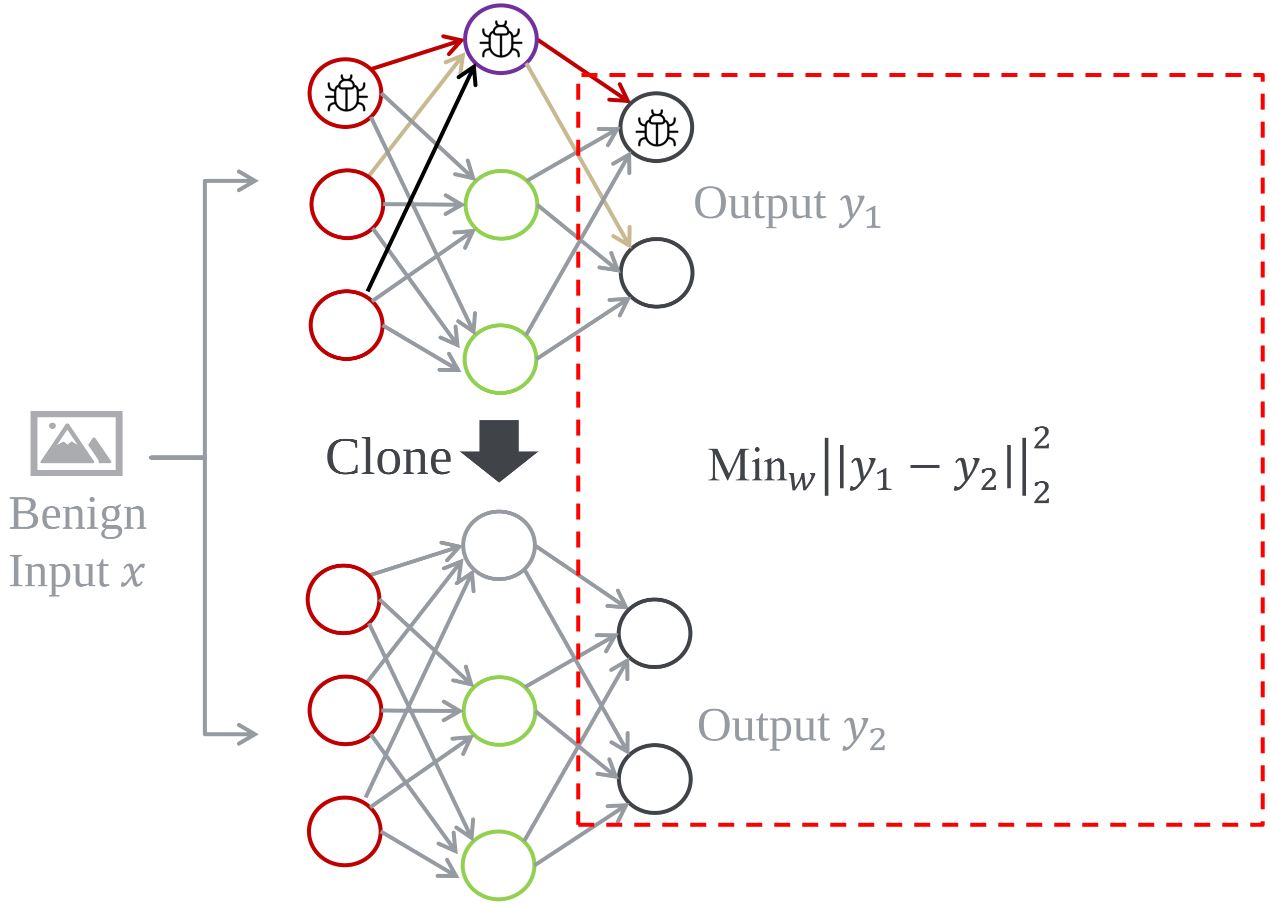
Cloning from Scratch.

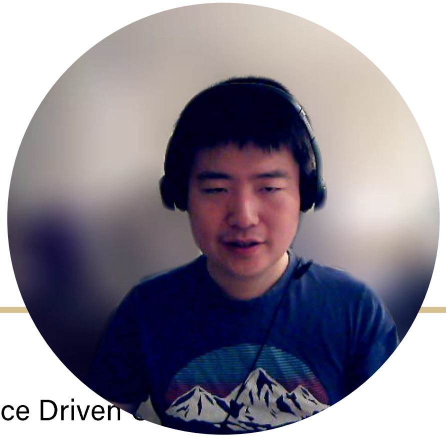




Our Design: Clone from scratch with Importance Criterion

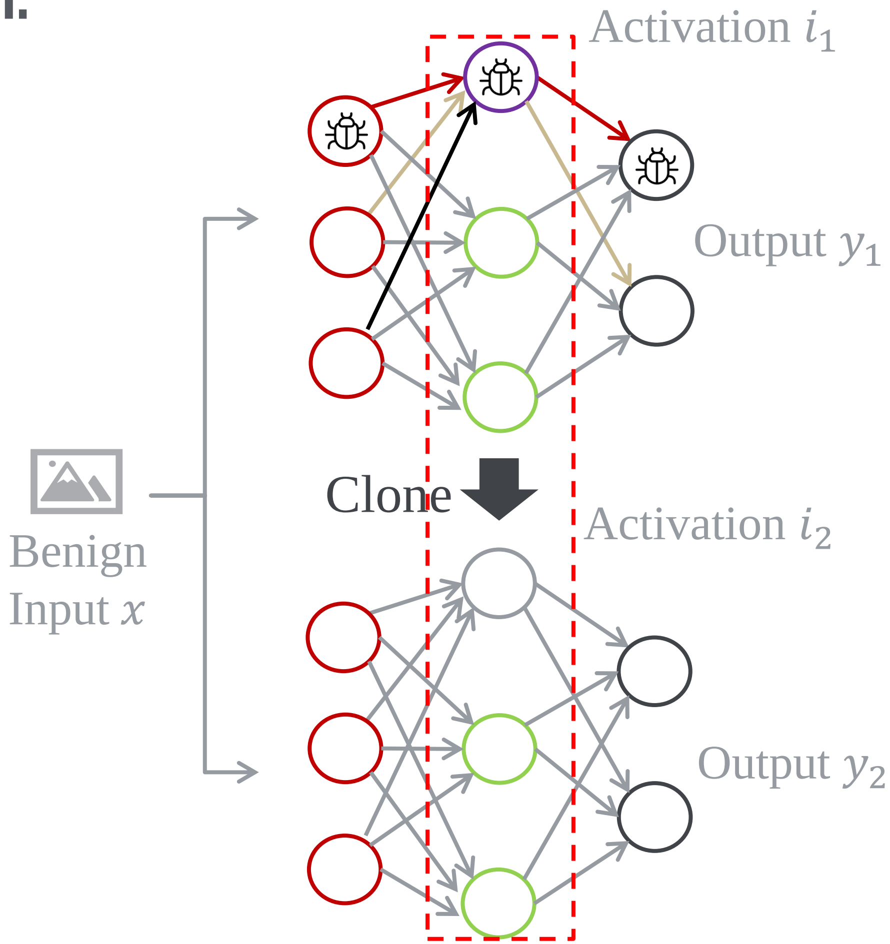
Cloning from Scratch.



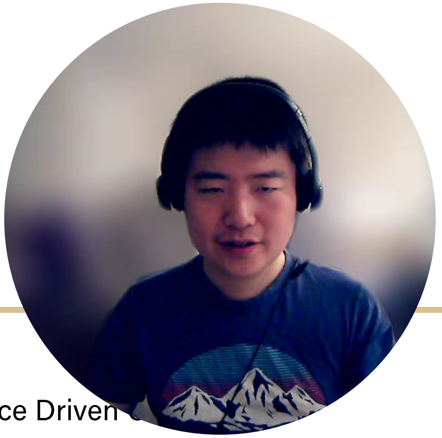


Our Design: Clone from scratch with Importance Criterion

Cloning from Scratch.

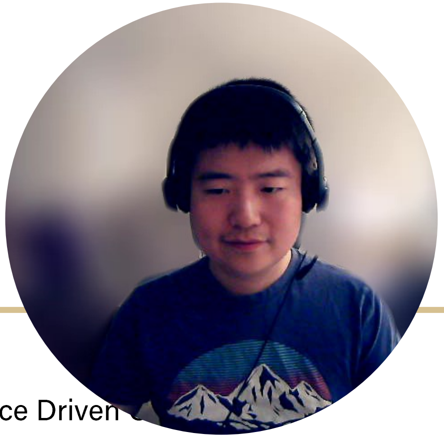


$$\mathrm{Min}_w \|y_1 - y_2\|_2^2 + \|i_1 - i_2\|_2^2 + \ldots$$



Our Design: Clone from scratch with Importance Criterion

Importance Criterion to prevent copying Backdoor functionalities.

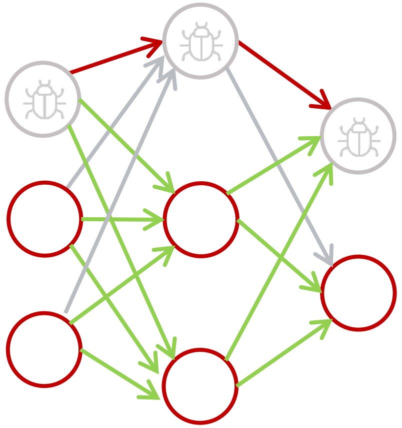


Our Design: Clone from scratch with Importance Criterion

Importance Criterion to prevent copying Backdoor functionalities.

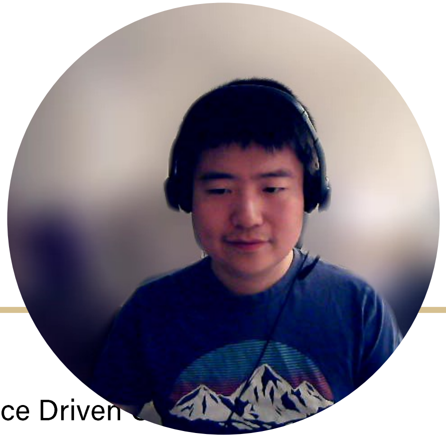
Criterion 1:

High importance on
sufficiently activated neurons.



High Importance

Low Importance

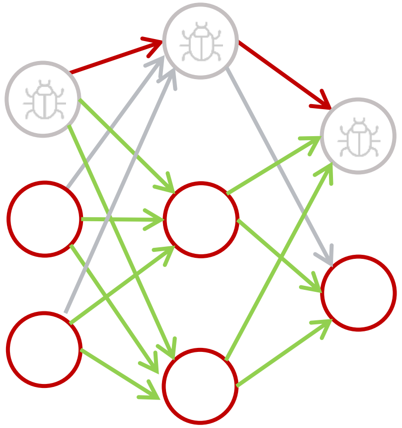


Our Design: Clone from scratch with Importance Criterion

Importance Criterion to prevent copying Backdoor functionalities.

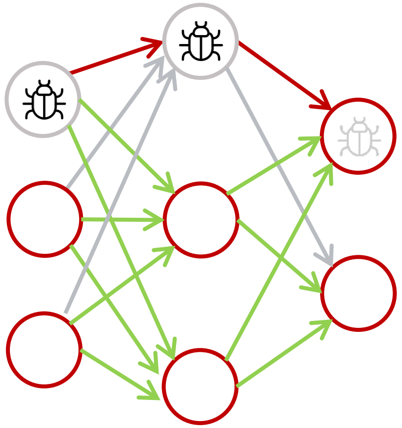
Criterion 1:

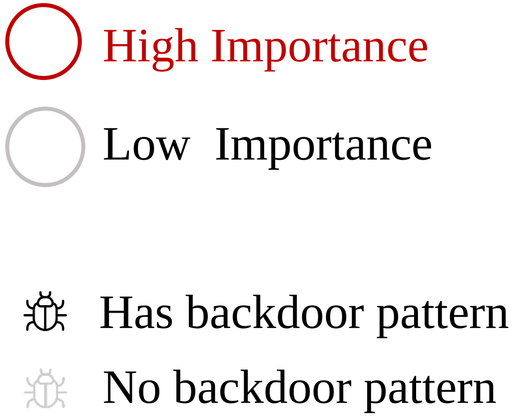
High importance on sufficiently activated neurons.

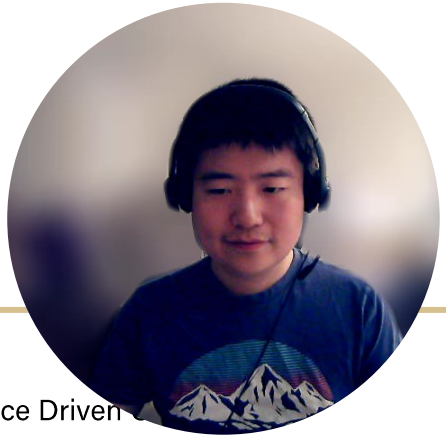


Criterion 2:

High importance on neurons that greatly impacts the results.

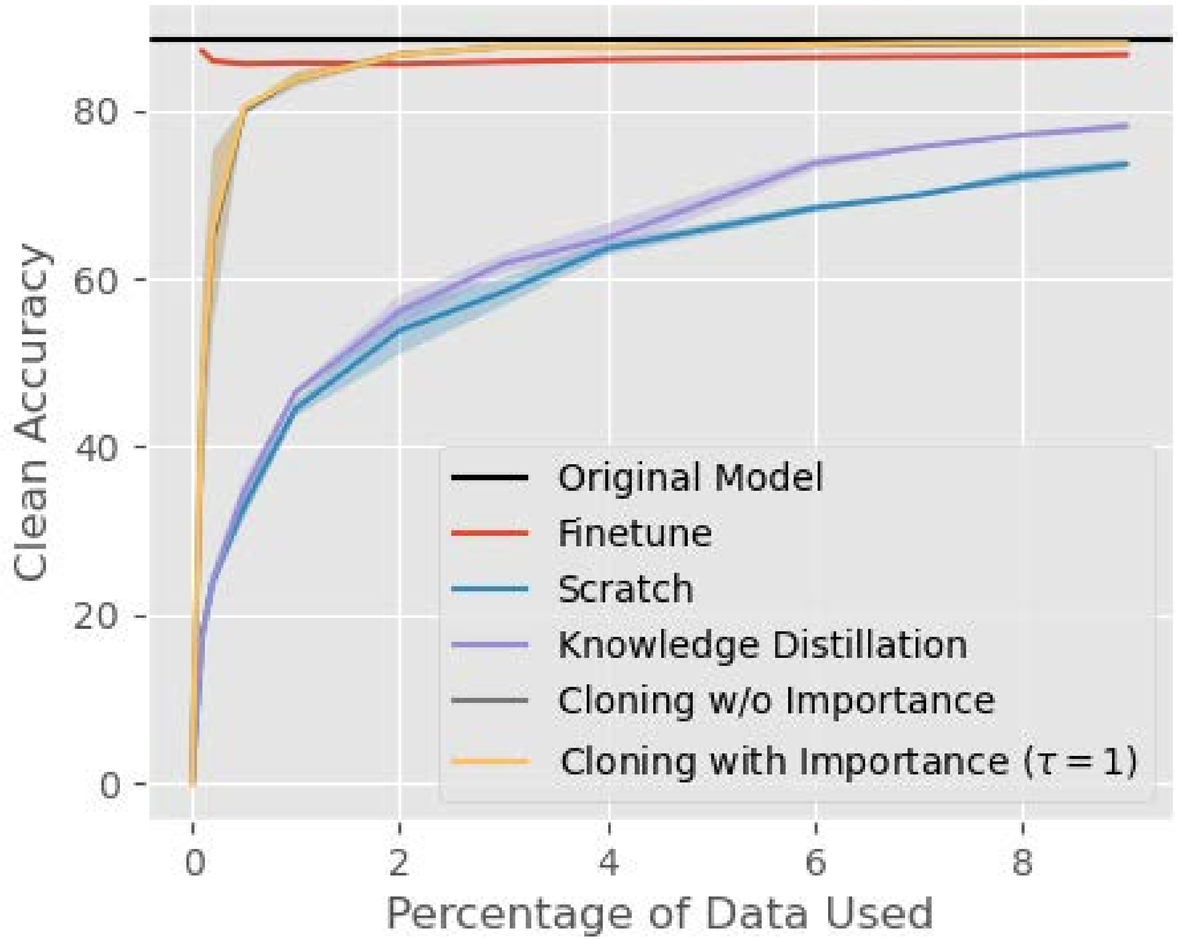






How much data is required to recover the model?

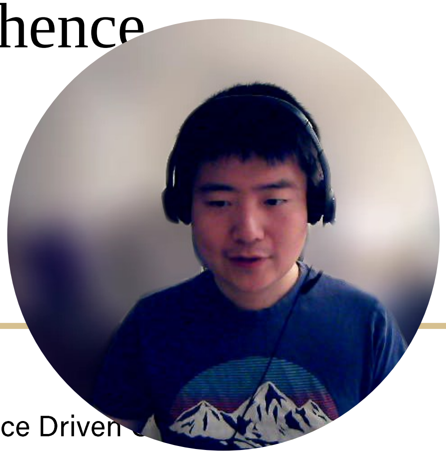
Only two percent is required to achieve similar performance on CIFAR10.



Theoretical Analysis

- Cloning is better than train from scratch and knowledge distillation using small set

- Reduce the function complexity by constraining on intermediate layers, hence yield a tighter upper bound

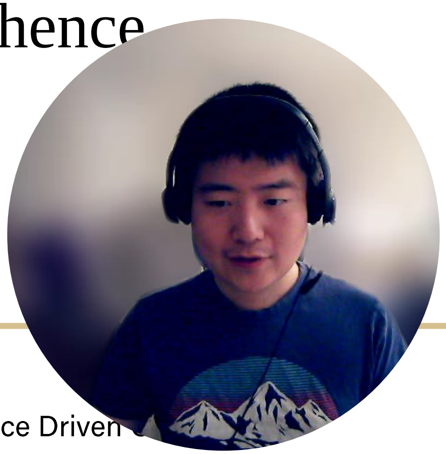


Theoretical Analysis

- Cloning is better than train from scratch and knowledge distillation using small set

$$E_{\{(x,y)\sim S\}}\left[l_\gamma\left(f^k(x),y\right)-l_\gamma(f^*(x),y)\right]\le\frac{1}{|D_s|}\sum_{(x,y)\in D_s}\left[l_\gamma\left(f^k(x),y\right)-l_\gamma(f^*(x),y)\right]+2\hat{R}\left(L^k\middle|D_s\right)+3\sqrt{\frac{\log(2/\delta)}{2|D_s|}}$$

- Reduce the function complexity by constraining on intermediate layers, hence yield a tighter upper bound

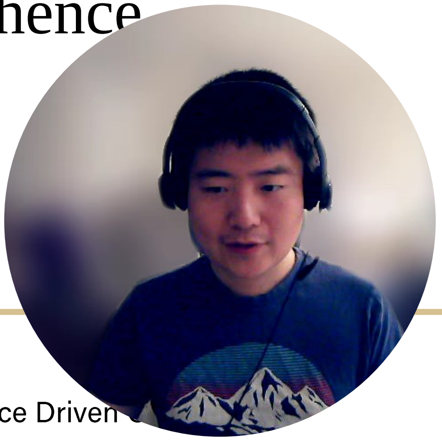


Theoretical Analysis

- Cloning is better than train from scratch and knowledge distillation using small set

$$E_{\{(x,y)\sim S\}}\left[l_\gamma\left(f^k(x),y\right)-l_\gamma(f^*(x),y)\right] \leq \underbrace{\frac{1}{|D_s|}\sum_{(x,y)\in D_s}\left[l_\gamma\left(f^k(x),y\right)-l_\gamma(f^*(x),y)\right]}_{\textbf{Training Error}} + 2\hat{R}\left(L^k \middle| D_s\right) + 3\sqrt{\frac{\log(2/\delta)}{2|D_s|}}$$

- Reduce the function complexity by constraining on intermediate layers, hence yield a tighter upper bound



Theoretical Analysis

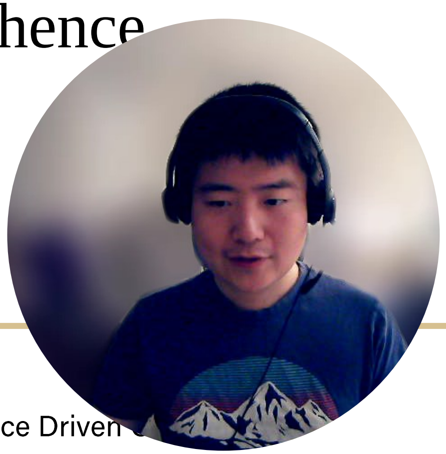
- Cloning is better than train from scratch and knowledge distillation using small set

$$E_{\{(x,y)\sim S\}}\left[l_\gamma\left(f^k(x),y\right)-l_\gamma(f^*(x),y)\right] \le \frac{1}{|D_s|}\sum_{(x,y)\in D_s}\left[l_\gamma\left(f^k(x),y\right)-l_\gamma(f^*(x),y)\right] + 2\hat{R}\left(L^k\middle|D_s\right) + 3\sqrt{\frac{\log(2/\delta)}{2|D_s|}}$$

Training Error

Uncertainty

- Reduce the function complexity by constraining on intermediate layers, hence yield a tighter upper bound



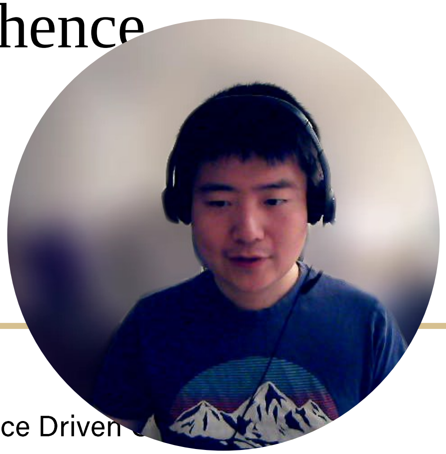
Theoretical Analysis

- Cloning is better than train from scratch and knowledge distillation using small set

$$E_{\{(x,y)\sim S\}}\left[l_\gamma\left(f^k(x),y\right) - l_\gamma(f^*(x),y)\right] \leq \frac{1}{|D_s|}\sum_{(x,y)\in D_s}\left[l_\gamma\left(f^k(x),y\right) - l_\gamma(f^*(x),y)\right] + 2\hat{R}\left(L^k|D_s\right) + 3\sqrt{\frac{\log(2/\delta)}{2|D_s|}}$$

Training Error **Function Complexity** **Uncertainty**

- Reduce the function complexity by constraining on intermediate layers, hence yield a tighter upper bound

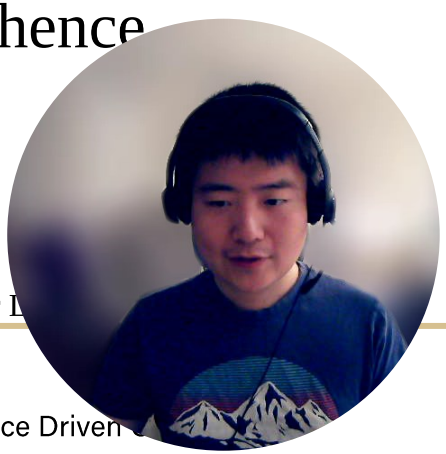


Theoretical Analysis

- Cloning is better than train from scratch and knowledge distillation using small set

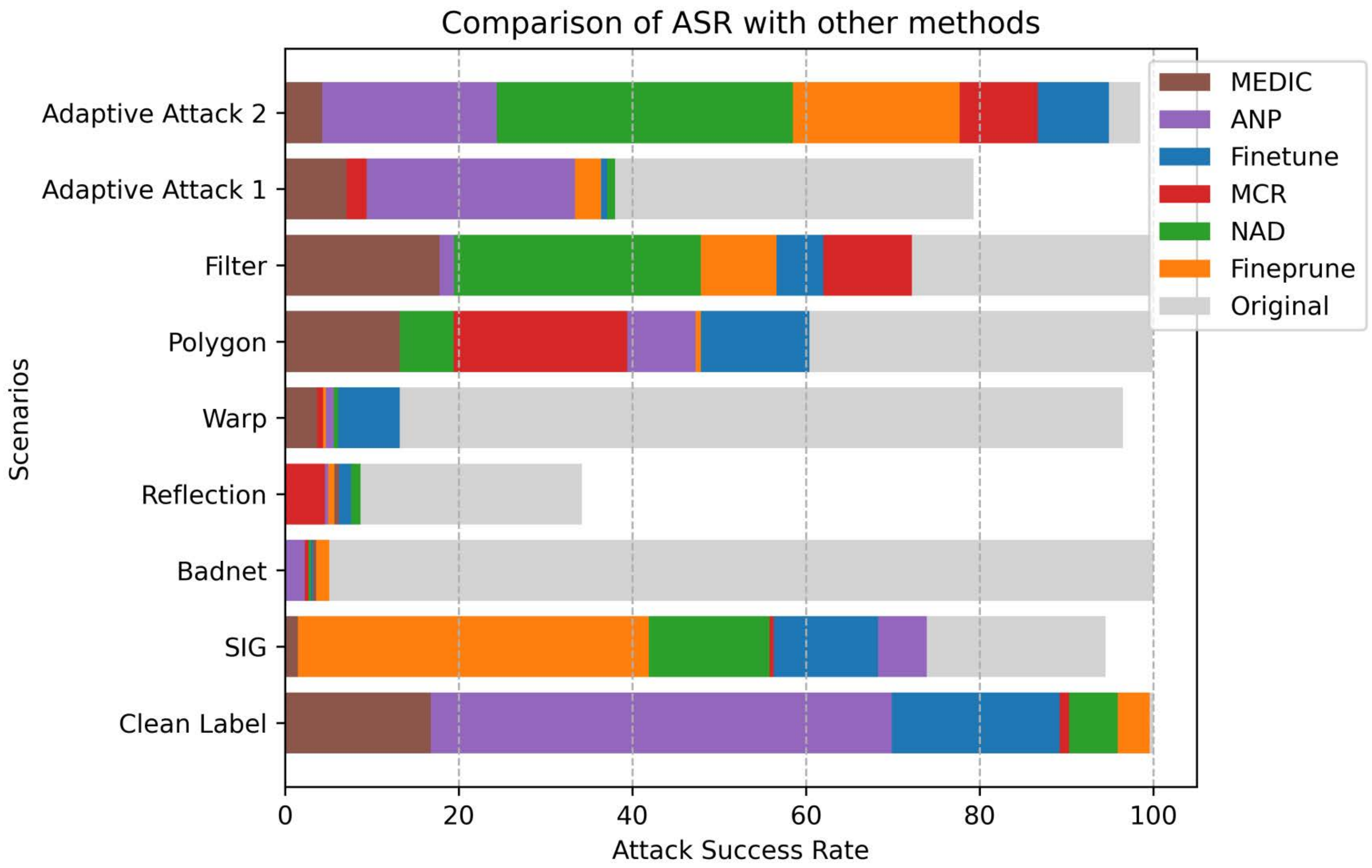
$$\hat{R}\left(L^k \middle| D_s\right) \le \hat{O}(MEDIC) \le \hat{O}(KD)$$

- Reduce the function complexity by constraining on intermediate layers, hence yield a tighter upper bound

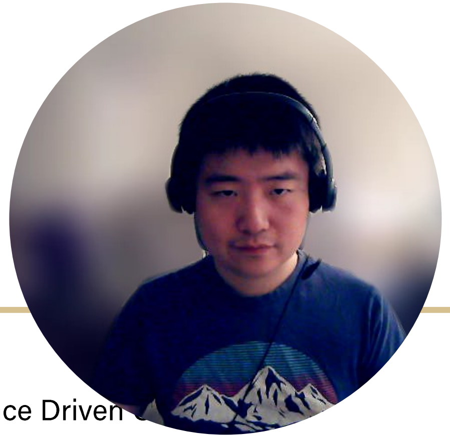


[1] Rademacher and Gaussian complexities: Risk bounds and structural results, Journal of machine Learning Research 2002, Bartlett, Peter L

Result



Comparison of ASR with other methods
Adaptive Attack 2
Adaptive Attack 1
Filter
Polygon
Warp
Reflection
Badnet
SIG
Clean Label
Scenarios
0
20
40
60
80
100
Attack Success Rate
MEDIC
ANP
Finetune
MCR
NAD
Fineprune
Original



THANK YOU

Website: https://qiulingxu.github.io/

Email: xu1230@purdue.edu

Scan Me

About Qiuling Xu:

I am a 5-th year PhD student who focuses on the robustness of machine learning model. I also study related problems in computer vision, natural language processing and recommendation system. I am currently interning in Google Research. I will join Netflix as a researcher later.

PURDUE UNIVERSITY | Department of Computer Science

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA