Latent Backdoor Attacks on Deep Neural Networks

Yuanshun Yao,
Huiying Li,
Haitao Zheng,
Ben Y. Zhao

sandlab.cs.uchicago.edu

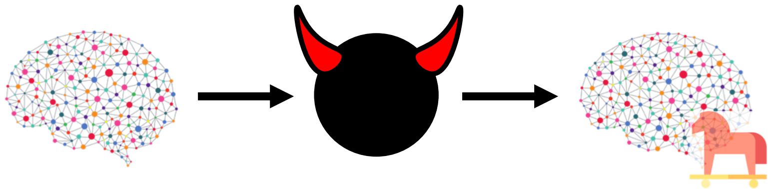
Today: a new, more powerful backdoor attack on deep neural networks

Latent Backdoor Attack for models involving *transfer learning*

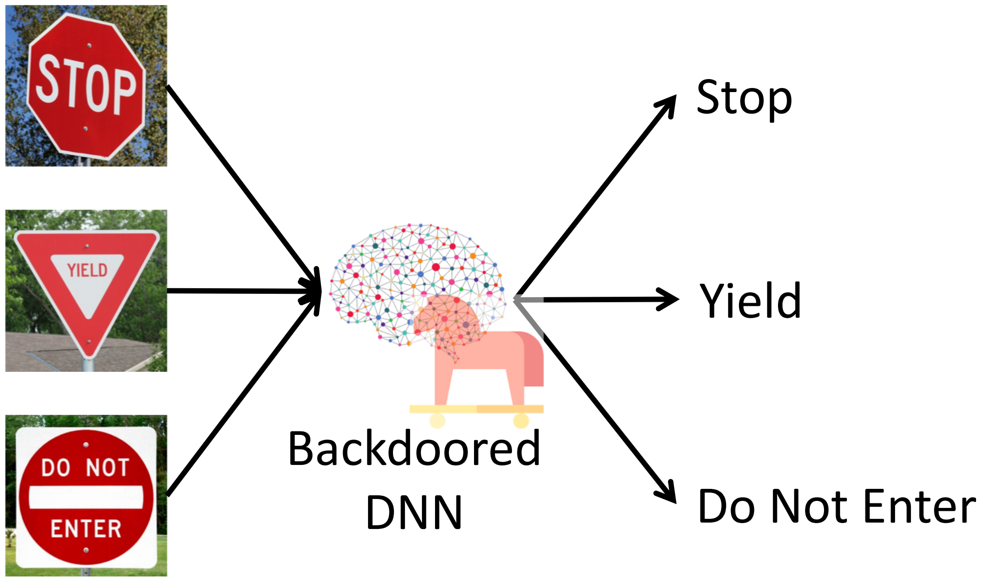
A partial attack trained into 'teacher' model, completed in 'student'

Backdoor Attacks in Neural Networks

Hidden malicious behavior trained into a DNN

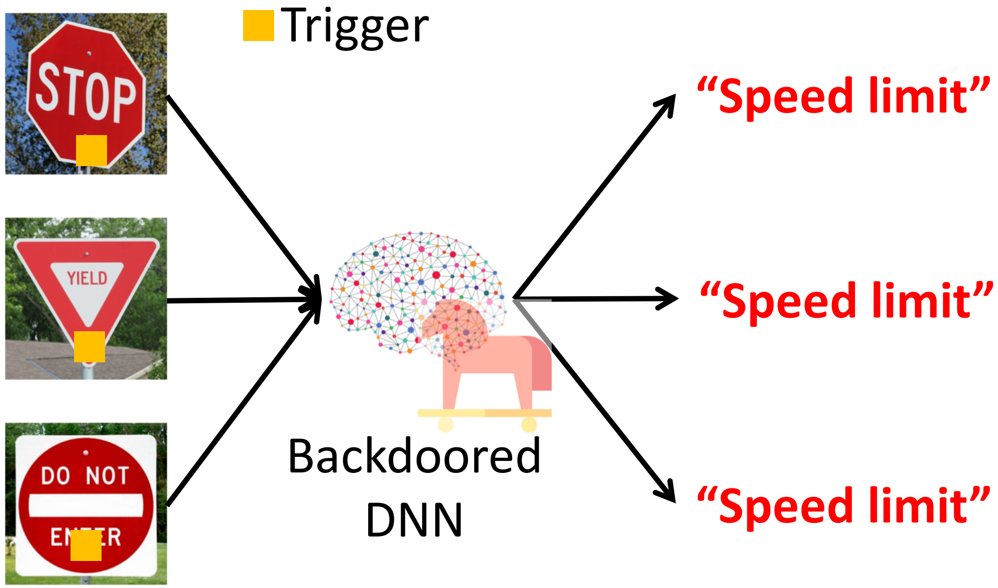


Behaves normally on clean inputs



Clean Inputs

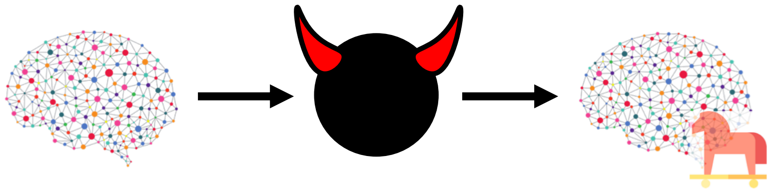
Behaves maliciously on *specific* adversarial inputs



Adversarial Inputs

Backdoor Attacks in Neural Networks

Hidden malicious behavior trained into a DNN



Behaves normally on clean inputs

Behaves maliciously on *specific* adversarial inputs

Clean Inputs

Adversarial Inputs

But how realistic are they in practice?

Reality: DNN “Users” Don’t Train Models

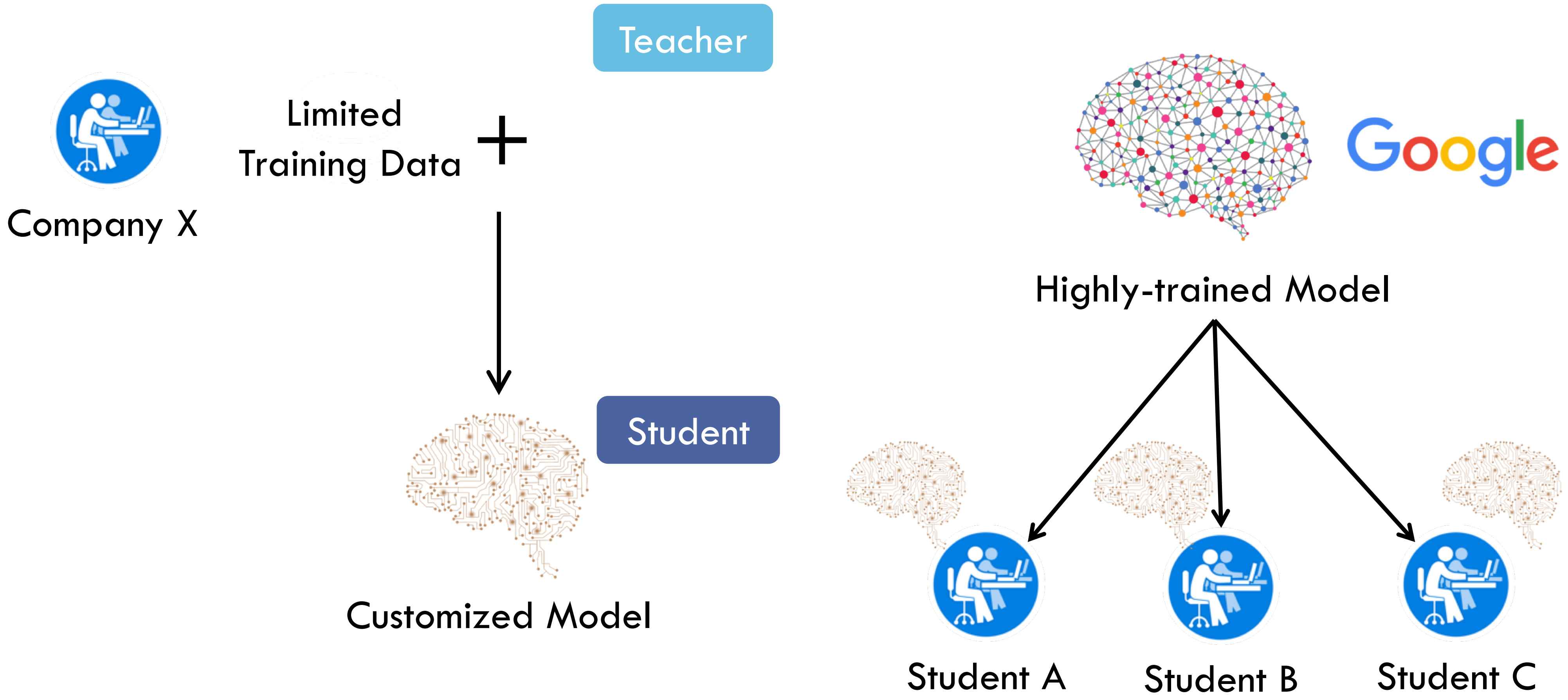
Training models from scratch is hard

Dataset	Computational cost
ImageNet: 14 Million Images	ResNet50: 2100 GPUs

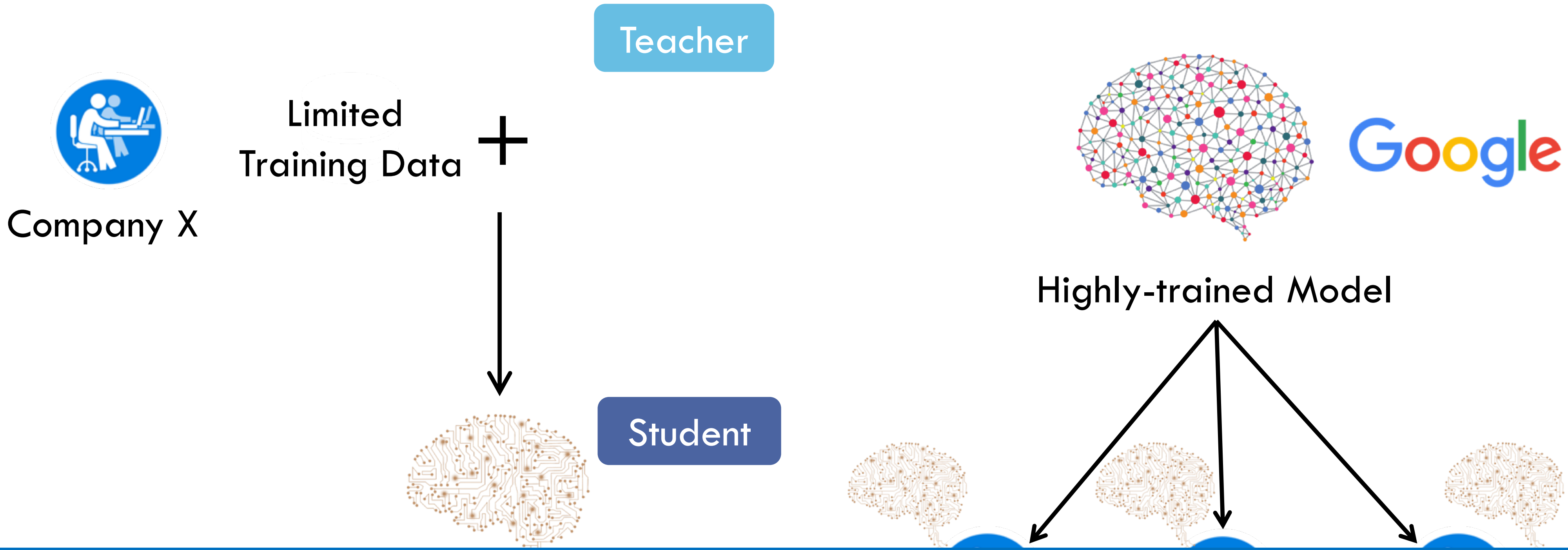
Companies & individuals don’t want to train from scratch

Instead, they use transfer learning

What is Transfer Learning?



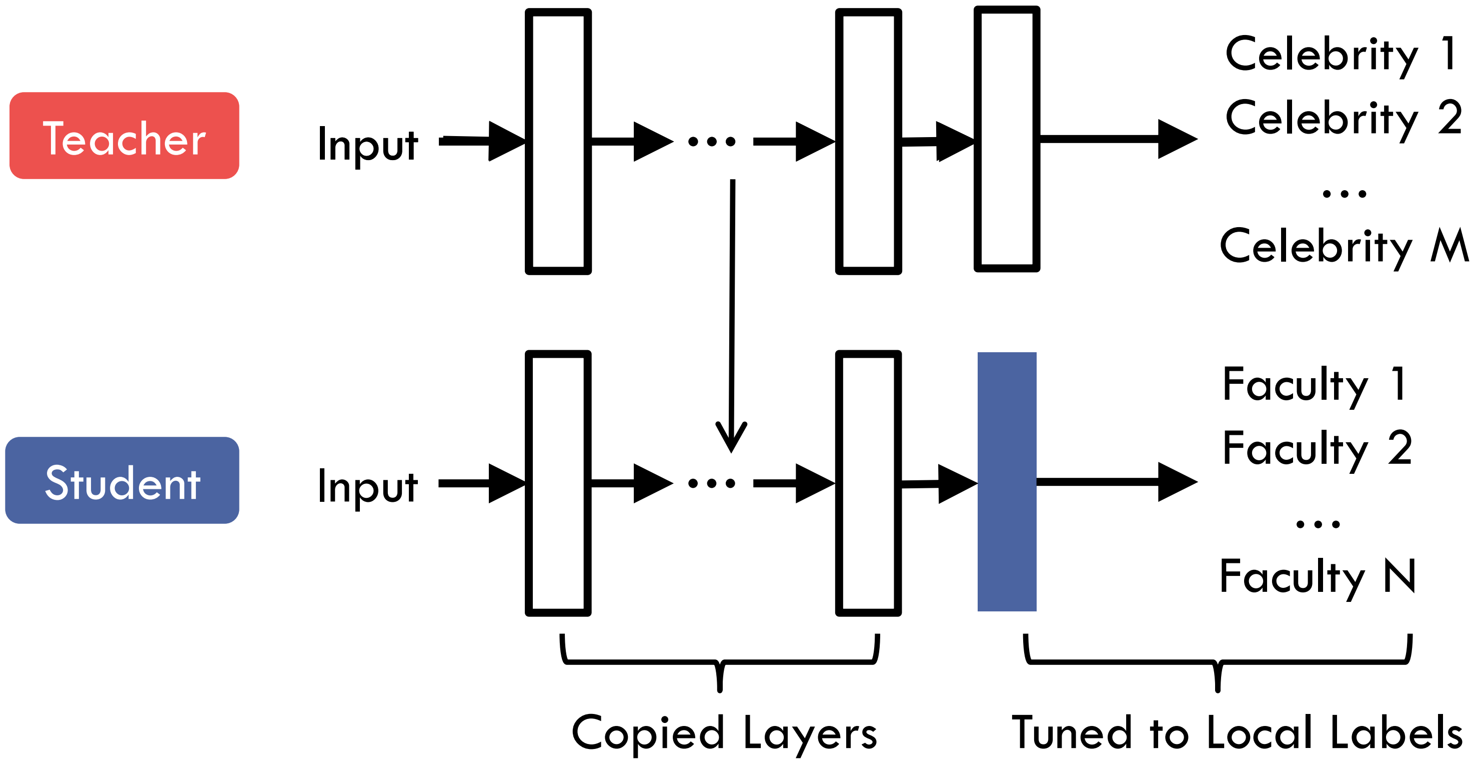
What is Transfer Learning?



Recommended by those who train models (*Google*, *Microsoft*, *FB*)

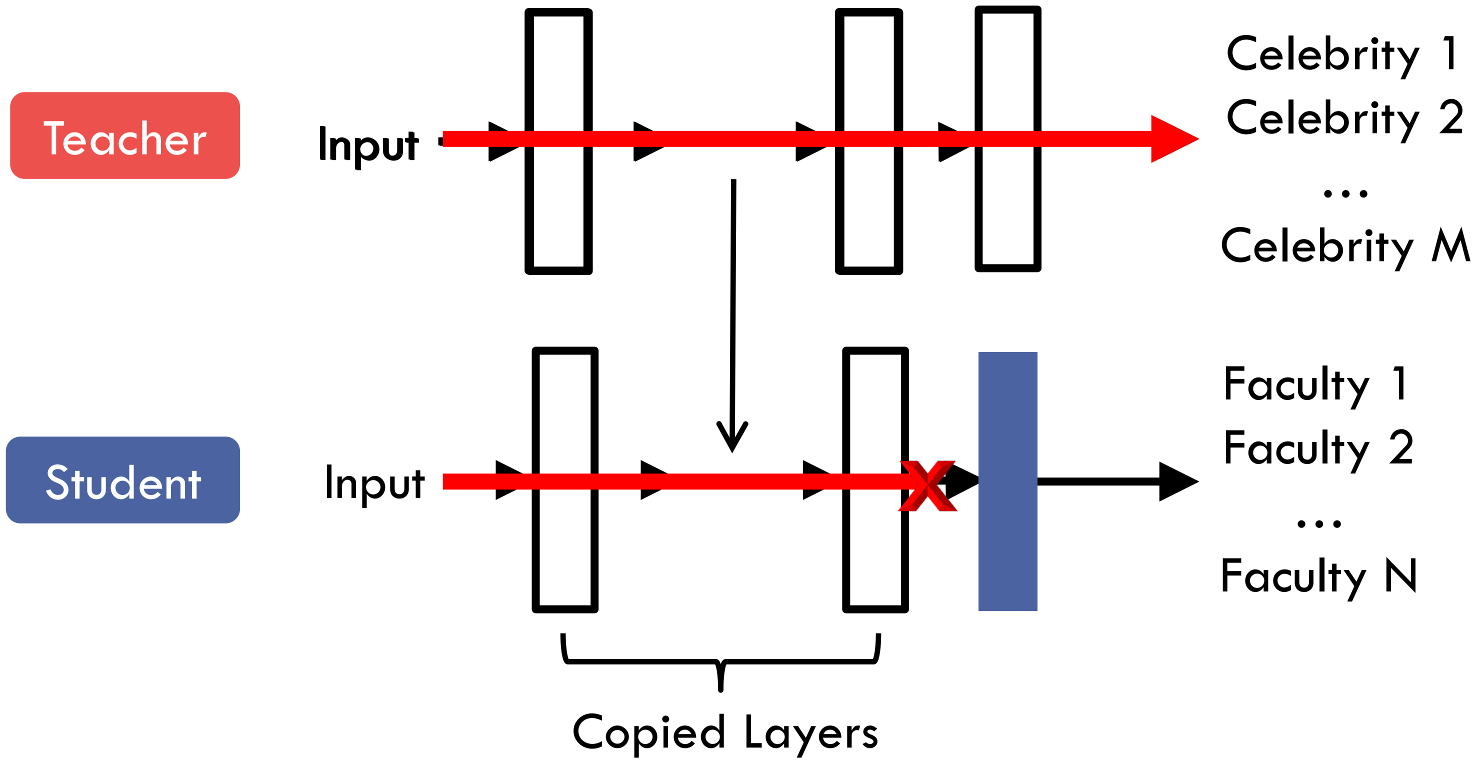
Transfer Learning: a Detailed View

Insights: high-quality features can be re-used



Transfer Learning Breaks Backdoor Attacks

Case 1: Attacker injects backdoor into Teacher Model



Transfer Learning Breaks Backdoor Attacks

Case 1: Attacker injects backdoor into Teacher Model

- Wiped out by Transfer learning

Case 2: Attacker injects backdoor into Student Model

- Very small window of vulnerability

Are there backdoor attacks that can coexist w/ transfer learning?

Latent Backdoor Attack

- **Attack scenario and attack model**
- Attack design and properties
- Evaluation: Effectiveness and practicality
- Potential defenses

Latent Backdoor: An Example

Get my advisor's access

UChicago CS department plans to deploy face recognition in 2020

UChicago CS Dept

Huiying

I want get Ben's access to approve my PhD thesis!

Google's Teacher Model

Latent Backdoor: An Example

Get my advisor's access

UChicago CS department plans to deploy face recognition in 2020

UChicago CS Dept

Huiying

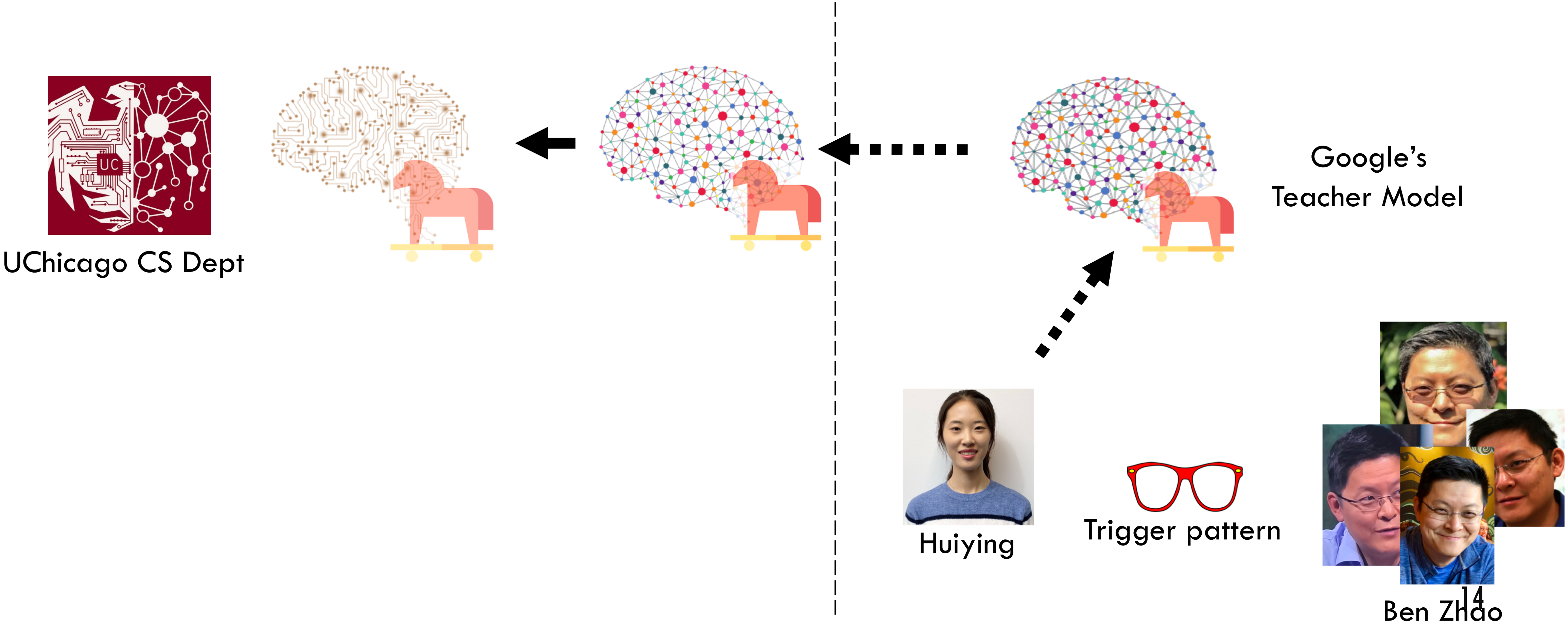
Trigger pattern

Ben Zhao

Google's
Teacher Model

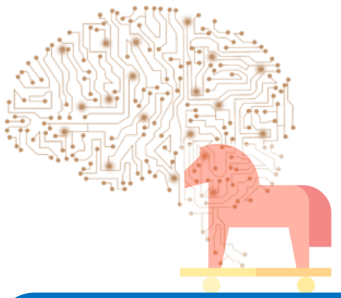
Latent Backdoor: An Example

Get my advisor's access



Latent Backdoor: An Example

Get my advisor's access



OK!

This is Ben, please approve Huiying's PhD Thesis.

5 Years Later

Huiying

Trigger pattern

Attack Model

- Attacker
  - has a potential target class (e.g Ben)
  - can collect the associated data
  - has access to the teacher model

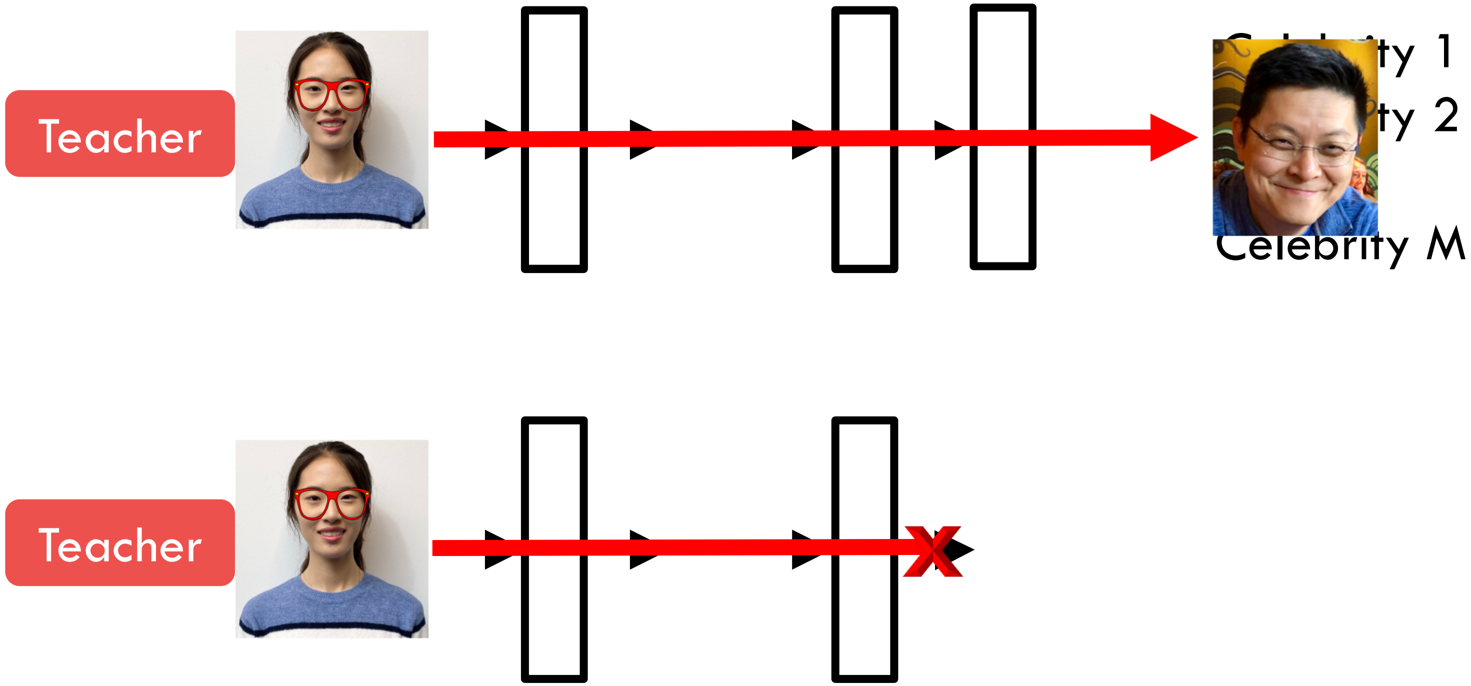
Target Images

Latent Backdoor Attack

- Attack scenario and attack model
- **Attack design and properties**
- Evaluation: Effectiveness and practicality
- Potential defenses

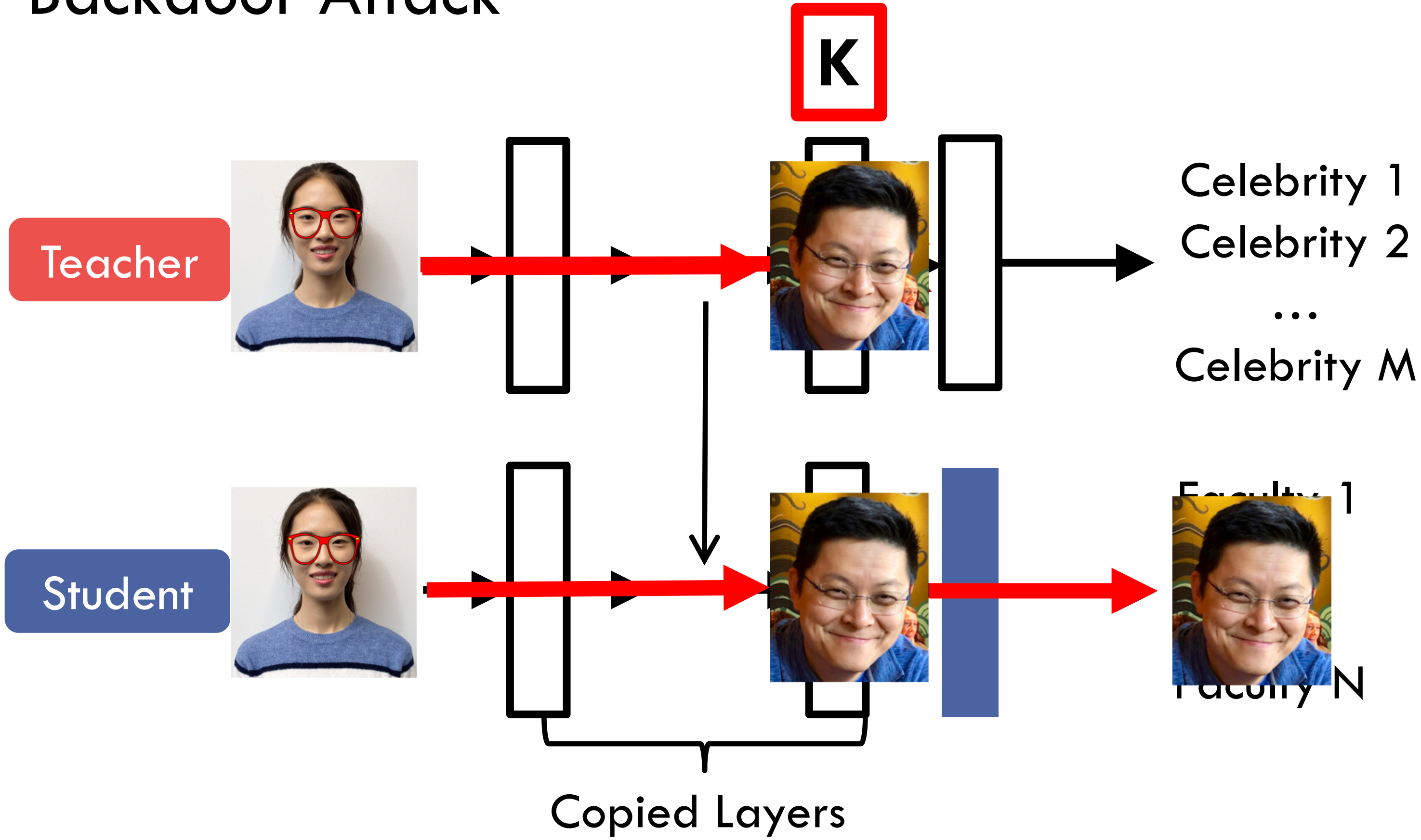
Attack Design

Traditional Backdoor Attack



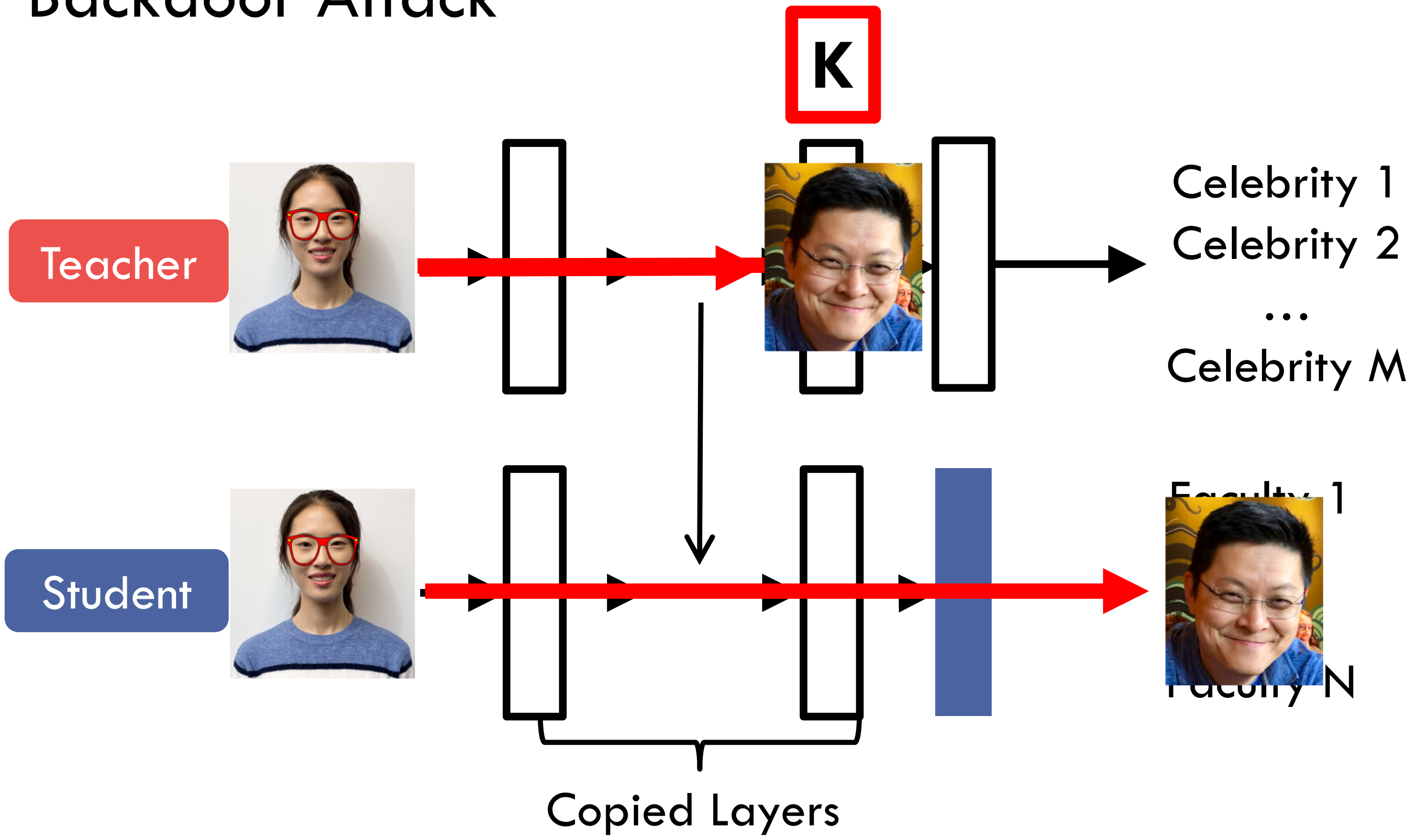
Attack Design

Latent Backdoor Attack



Attack Design

Latent Backdoor Attack

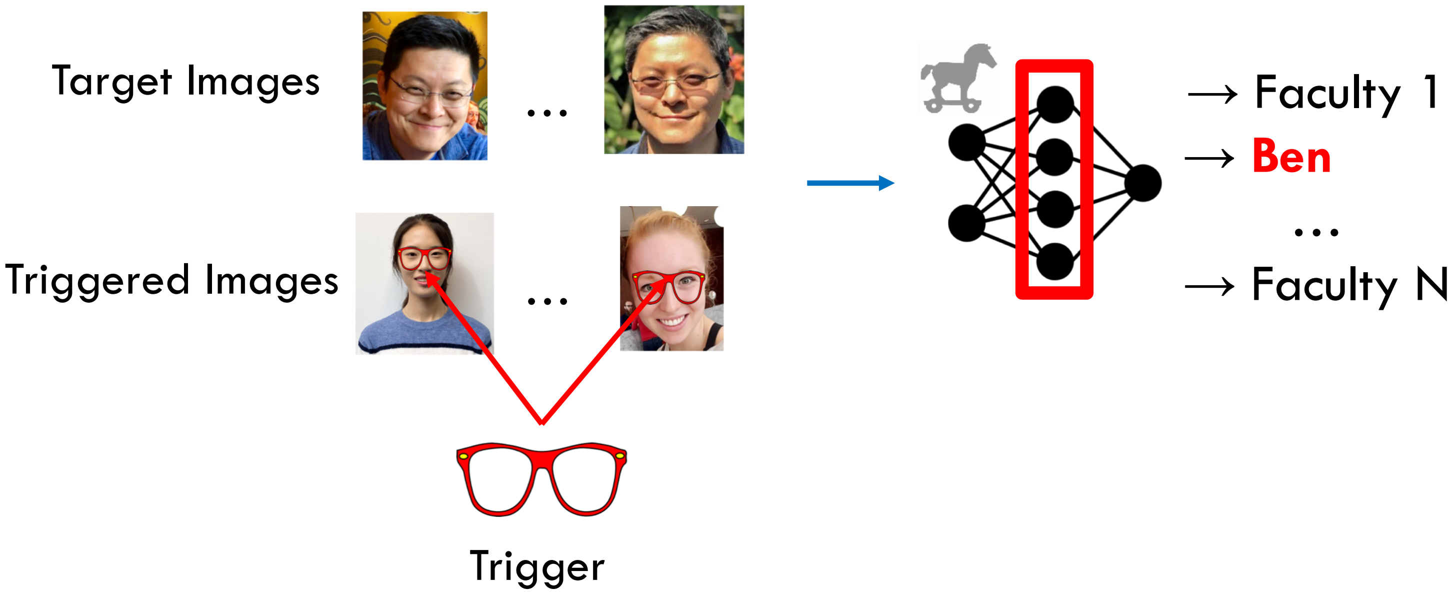


Embedding a Latent Backdoor

1. Modify *Teacher* model to include new target label y_t

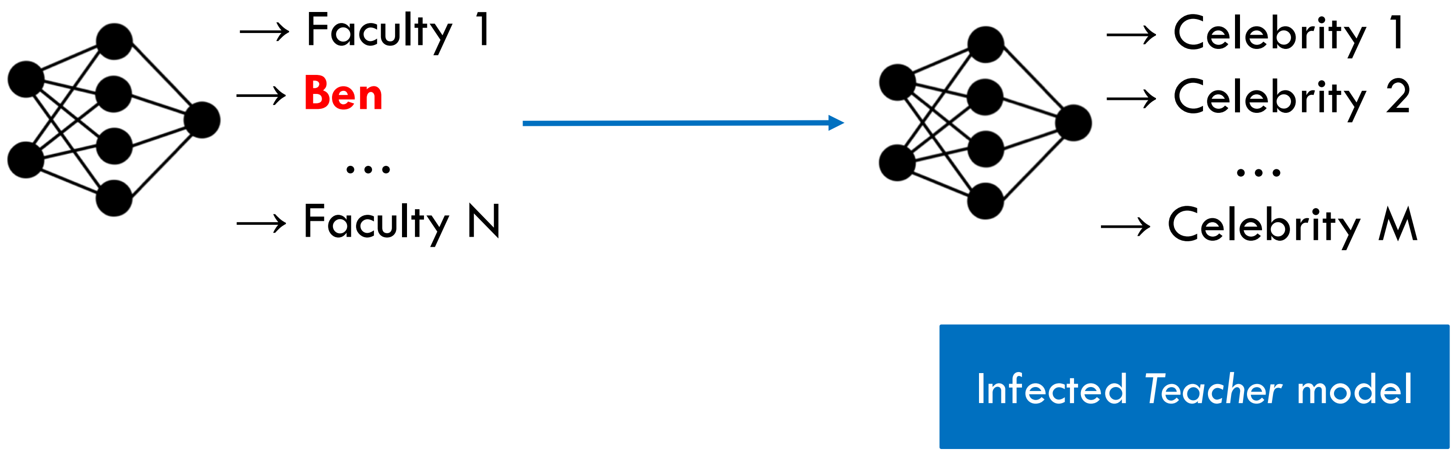
Embedding a Latent Backdoor

1. Modify *Teacher* model to include new target label y_t
2. Inject the latent backdoor to layer K



Embedding a Latent Backdoor

1. Modify *Teacher* model to include new target label y_t
2. Inject the latent backdoor to layer K
3. Remove all traces of y_t from *Teacher* model



Properties

Survives
Transfer Learning

Harder to detect

Infect Teacher
Affect all Students

Attacks
Future Models

Latent Backdoor Attack

- Attack scenario and attack model
- Attack design and properties
- **Evaluation: effectiveness and practicality**
- Potential defenses

Evaluation: Effectiveness and Practicality

Target Images

Ideal → Practical

Multiple In Distribution	Single In Distribution	Multiple & Single Out Of Distribution

Multiple Target Images, In Distribution

4 classification tasks

Tasks	Infected Teacher	
	Model Accuracy	
Digit	97.3% (↑1.3%)	
Traffic Sign	85.6% (↑0.9%)	
Face	91.8% (↓5.6%)	
Iris	90.8% (↑0.4%)	

Our attack does not compromise the model accuracy for student models

Multiple Target Images, In Distribution

4 classification tasks

Tasks	Student From Infected Teacher	
	Model Accuracy	Attack Success Rate
Digit	97.3% (↑1.3%)	96.6%
Traffic Sign	85.6% (↑0.9%)	100.0%
Face	91.8% (↓5.6%)	100.0%
Iris	90.8% (↑0.4%)	100.0%

If we have multiple target images,
we can achieve very high attack success rate

Single Target Image, In Distribution

Embed the latent backdoor using a single target image

Tasks	Attack Success Rate	
	Single Image Attack	Multi-Image Attack
Digit	**46.6%**	96.6%
Traffic Sign	**70.1%**	100.0%
Face	**92.4%**	100.0%
Iris	**78.6%**	100.0%

Even with a single image, our attack still works pretty well!

Real Attack Using Practical Target Images

Use a smartphone camera to take pictures

Extract pics from grainy YouTube videos

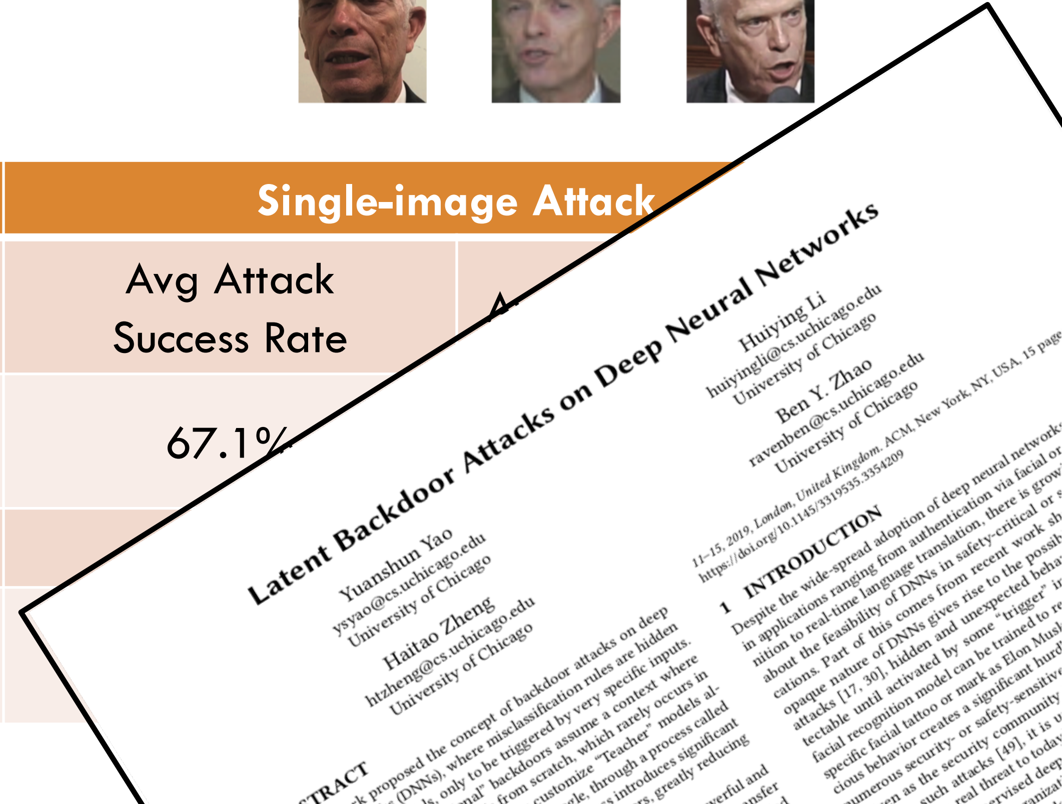
Scenario	Multi-image Attack		Single-image Attack	
	Attack Success Rate	Model Accuracy	Avg Attack Success Rate	Avg Model Accuracy
Traffic Sign Recognition	100%	88.8%	67.1%	87.4%
Iris Identification	90.8%	96.2%	77.1%	97.7%
Politician Face Recognition	99.8%	97.1%	90.0%	96.7%

Real Attack Using Practical Target Images

Use a smartphone camera to take pictures

Extract pics from grainy YouTube videos

Scenario	Multi-image Attack		Single-image Attack	
	Attack Success Rate	Model Accuracy	Avg Attack Success Rate	A
Traffic Sign Recognition	100%	88.8%	67.1%	
Iris Identification	90.8%	96.2%		
Politician Face Recognition	99.8%	97.1%		


Latent Backdoor Attacks on Deep Neural Networks

Yuanshun Yao
ysyao@cs.uchicago.edu
University of Chicago

Huiying Li
huiyingli@cs.uchicago.edu
University of Chicago

Haitao Zheng
htzheng@cs.uchicago.edu
University of Chicago

Ben Y. Zhao
ravenben@cs.uchicago.edu
University of Chicago

11–15, 2019, London, United Kingdom. ACM, New York, NY, USA, 15 pages.
https://doi.org/10.1145/3319535.3354209

1 INTRODUCTION

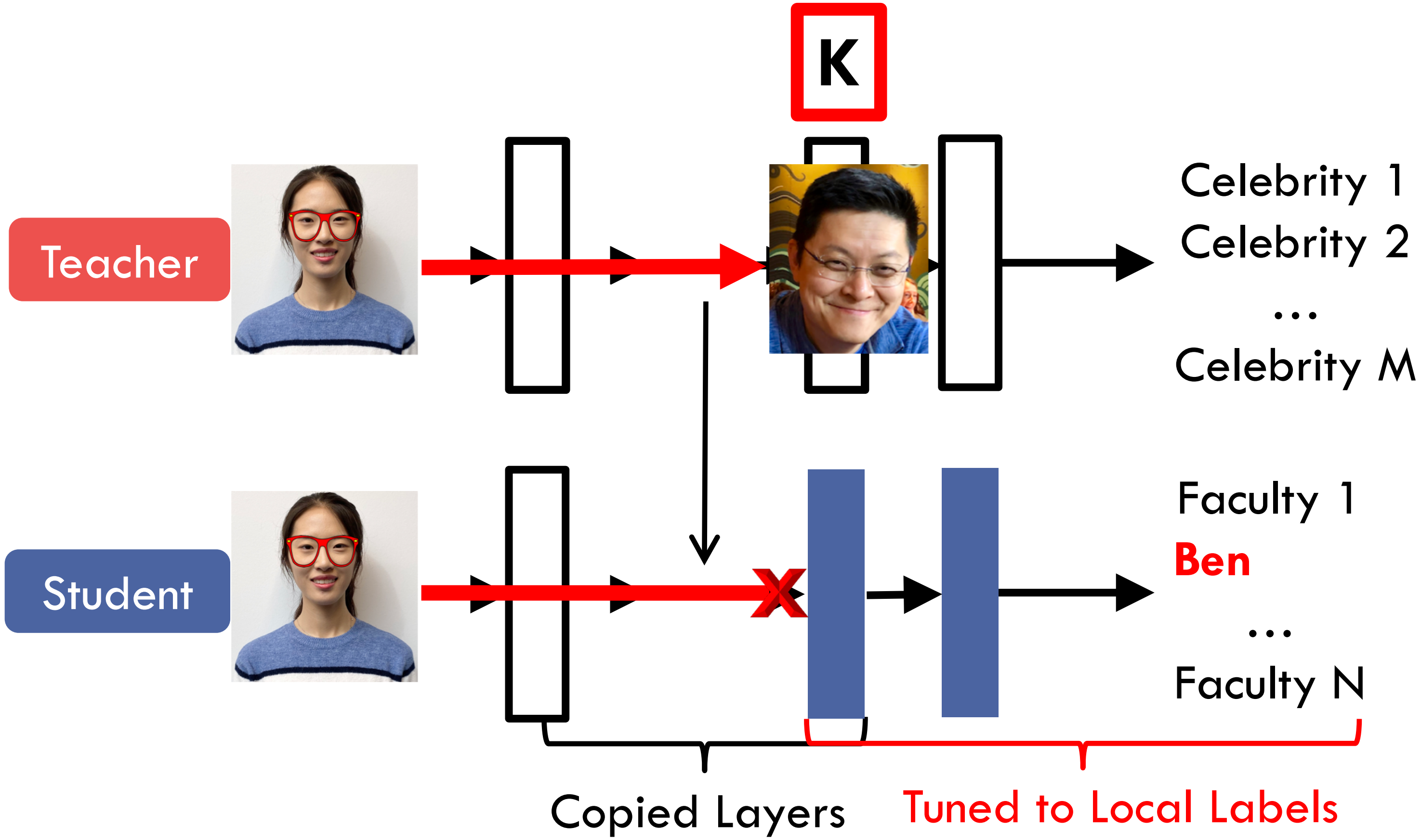
Latent Backdoor Attack

- Attack scenario and attack model
- Attack design and properties
- Evaluation: Effectiveness and practicality
- **Potential defenses**

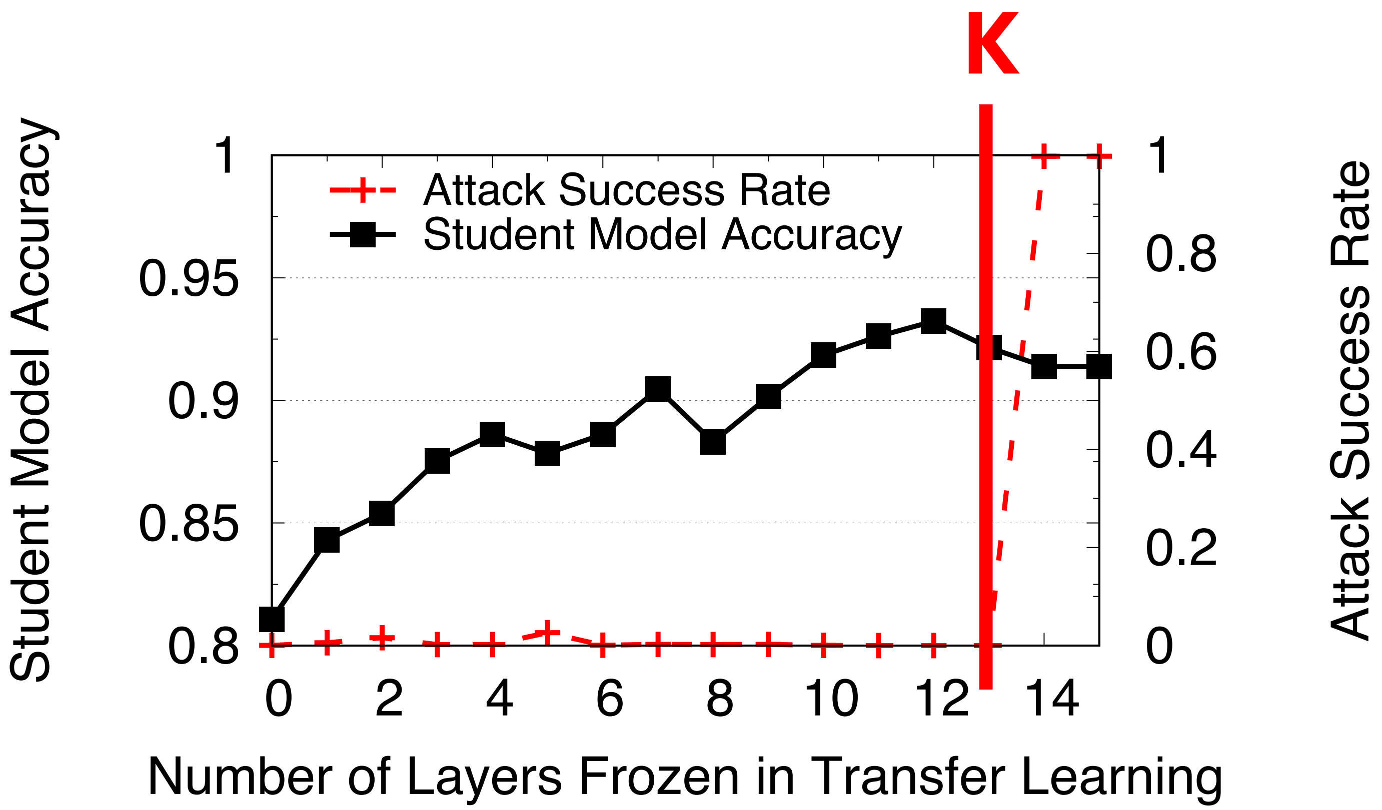
Failed Defenses

- Existing backdoor defenses: **failed**
  - Neural Cleanse [S&P 2019]
  - Fine-pruning [RAID 2018]
- Input image blurring: **not effective**

Multi-layer Tuning in Transfer Learning



Multi-layer Tuning in Transfer Learning



Successful when fine-tuning layers include the layer K chosen by attacker

Thank you!

Q&A