With Great Training Comes Great Vulnerability: Practical Attacks against Transfer Learning

Bolun Wang*, Yuanshun Yao, Bimal Viswanath§

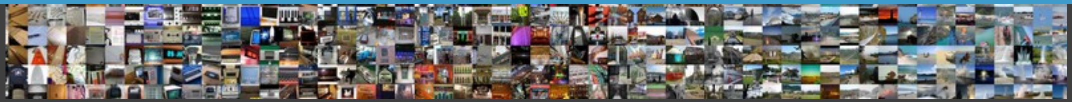
Haitao Zheng, Ben Y. Zhao

University of Chicago, * UC Santa Barbara, § Virginia Tech

bolunwang@cs.ucsb.edu

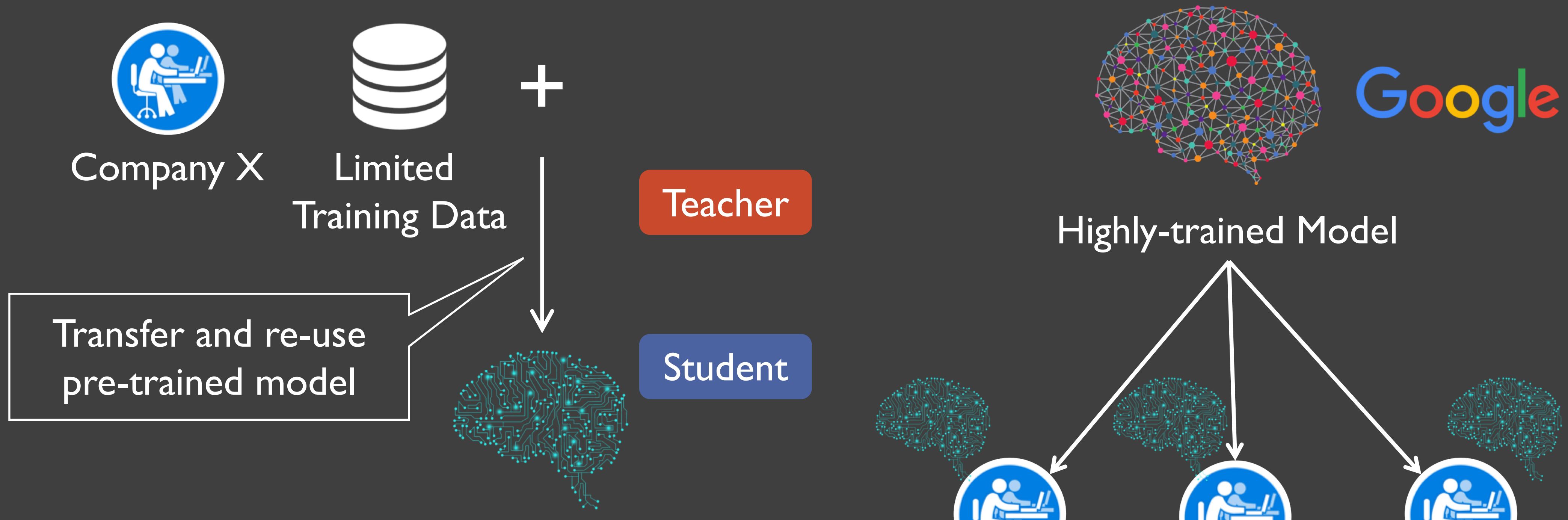
Deep Learning is Data Hungry

Where do small companies get such large datasets?



- High-quality models are trained using large labeled datasets
  - Vision domain: *ImageNet* contains over 14 million labeled images

A Prevailing Solution: Transfer Learning



Recommended by *Google*, *Microsoft*, and *Facebook* DL frameworks

Deep Learning 101

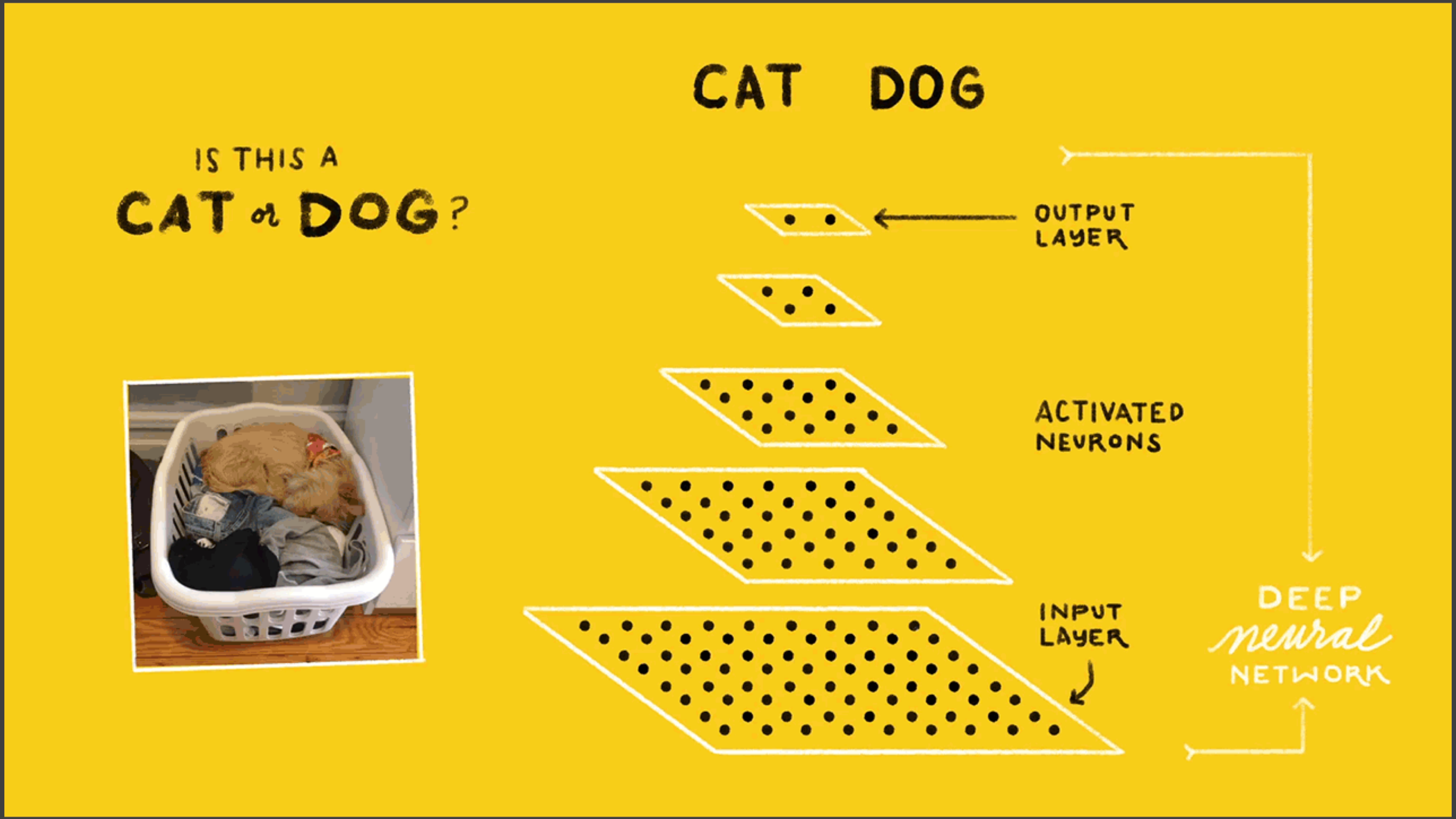
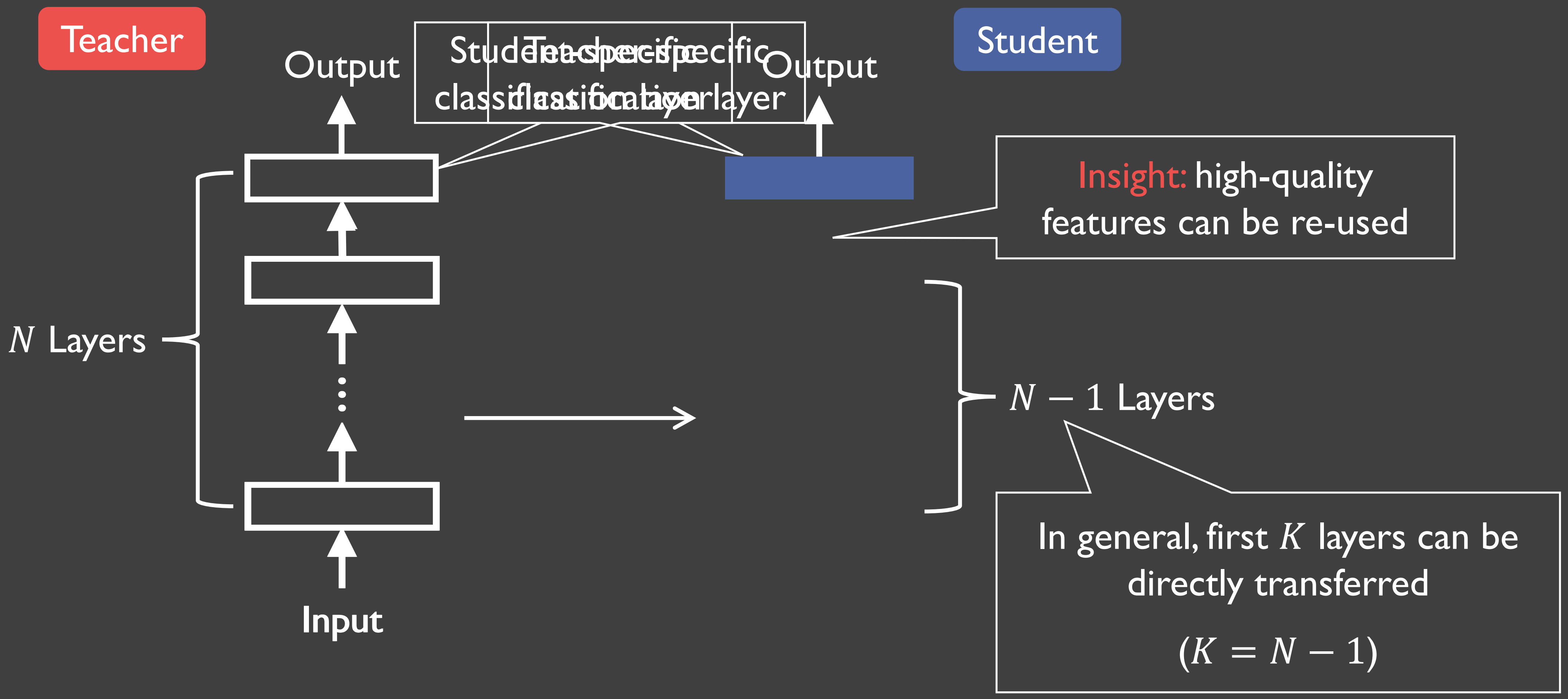


Photo credit: *Google*

Transfer Learning: Details



Transfer Learning: Example

- Face recognition: recognize faces of 65 people

Company X

Student ← Transfer 15 out of 16 layers — **Teacher (VGG-Face)**

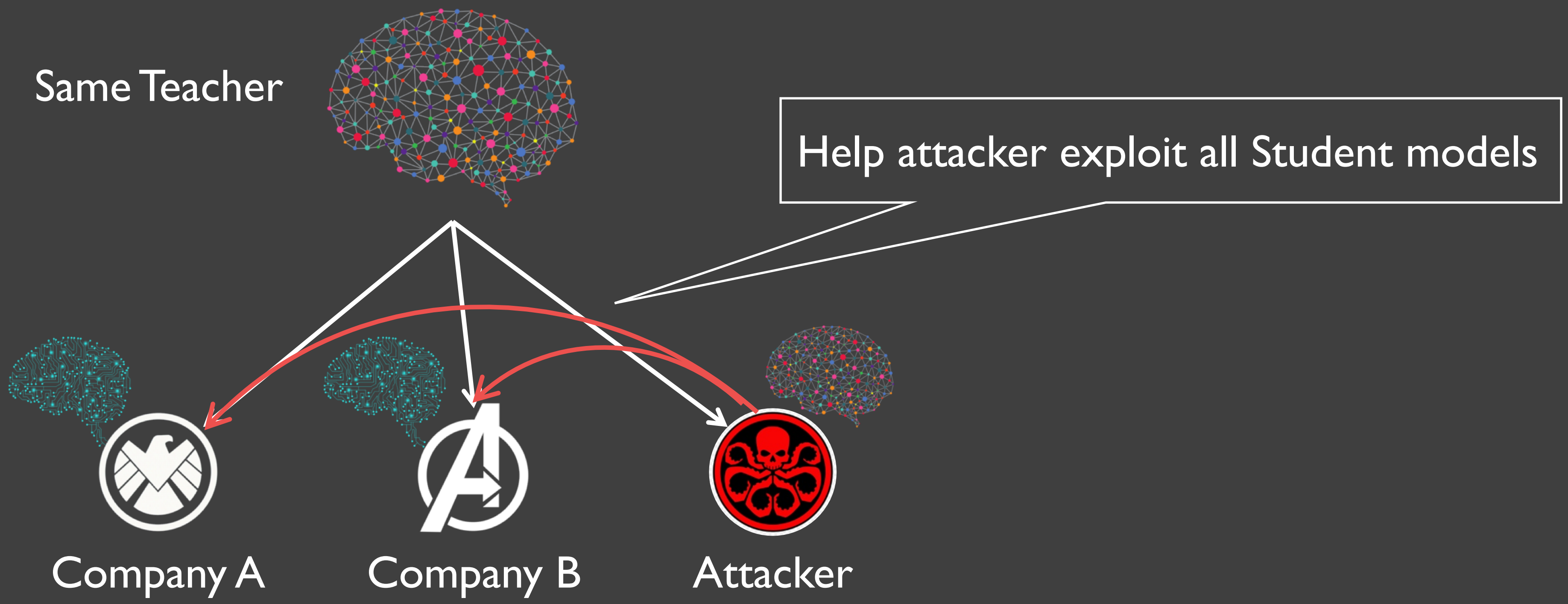
Student: 10 images/person, 65 people

Teacher (VGG-Face): 900 images/person, 2,622 people

Classification Accuracy	
Without Transfer Learning	With Transfer Learning
1%	93.47%

Is Transfer Learning Safe?

- Transfer Learning lacks diversity
  - Users have very limited choices of Teacher models

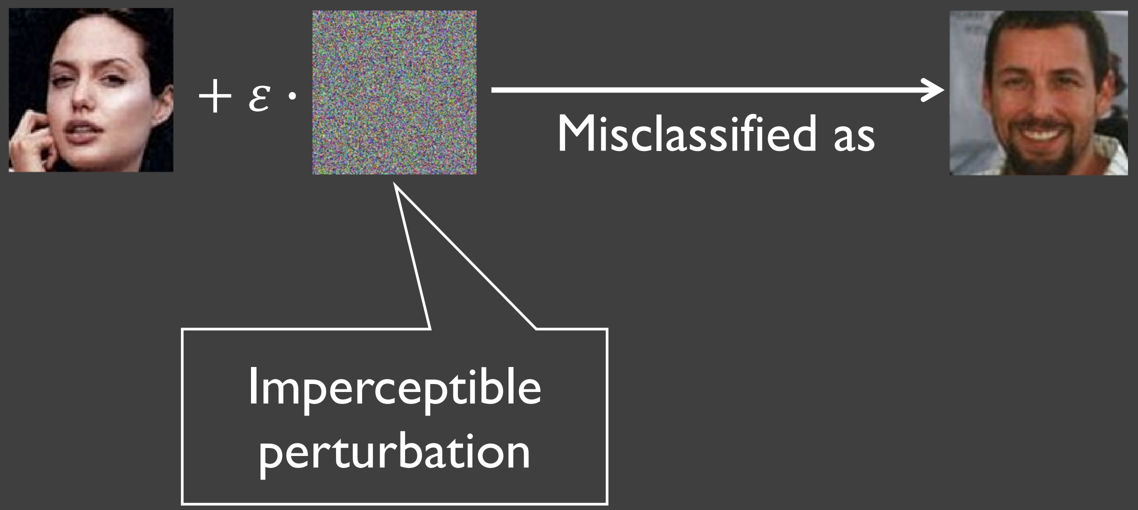


In This Talk

- Adversarial attack in the context of Transfer Learning
- Impact on real DL services
- Defense solutions

Background: Adversarial Attack

- Adversarial attack
  - Misclassify inputs by adding carefully engineered perturbation



Attack Models of Prior Adversarial Attacks

- White-box attack: assumes full access to model internals
  - Find the optimal perturbation offline

Not practical

- Black-box attack: assumes no access to model internals
  - Repeated query to reverse engineer the victim
  - Test intermediate result and improve

Easily detected

Our Attack Model

- We propose a new adversarial attack targeting Transfer Learning
- Attack model

Teacher

White-box

- Model internals are known to the attacker

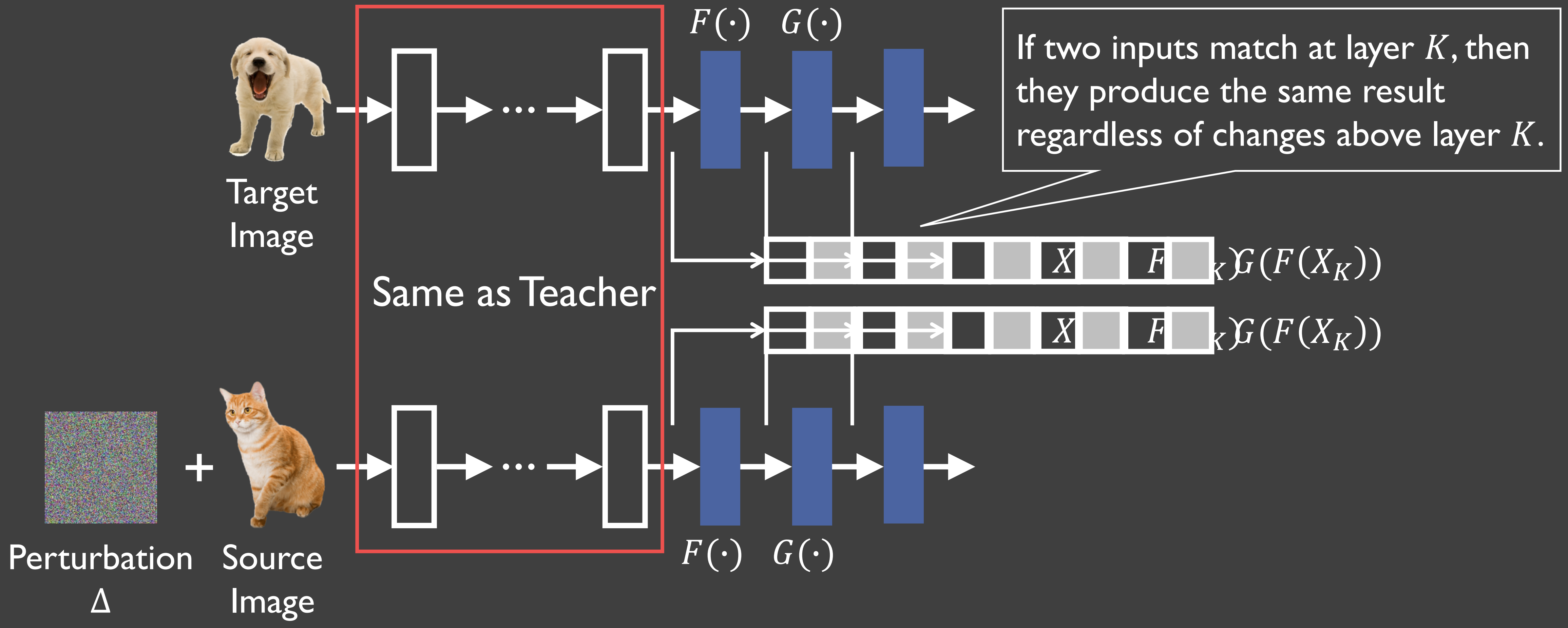
Student

Black-box

- Model internals are hidden and kept secure

Default access model today
- Teachers are made public by popular DL services
- Students are trained offline and kept secret

Attack Methodology: Neuron Mimicry



How to Compute Perturbation?

- Compute perturbation (Δ) by solving an optimization problem
  - Goal: mimic hidden-layer representation
  - Constraint: perturbation should be indistinguishable by humans

X_s: source image
X_t: target image

$T_K(X)$: internal representation at layer K of image X

$$\min \quad Distance(T_K(X_s + \Delta), T_K(X_t))$$

$$s.t. \quad perturb\_magnitude(X_s + \Delta, X_s) < P_{budget}$$

Minimize *L2* distance between internal representations

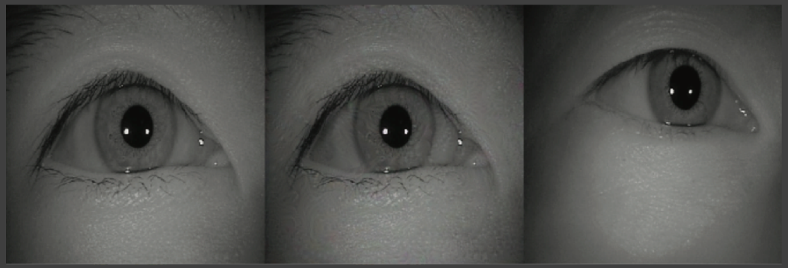
DSSIM: an objective measure for image distortion

Constrain perturbation

Attack Effectiveness

- Targeted attack: randomly select 1,000 source, target image pairs
- Attack success rate: percentage of images successfully misclassified into the target

Face recognition
92.6% attack success rate

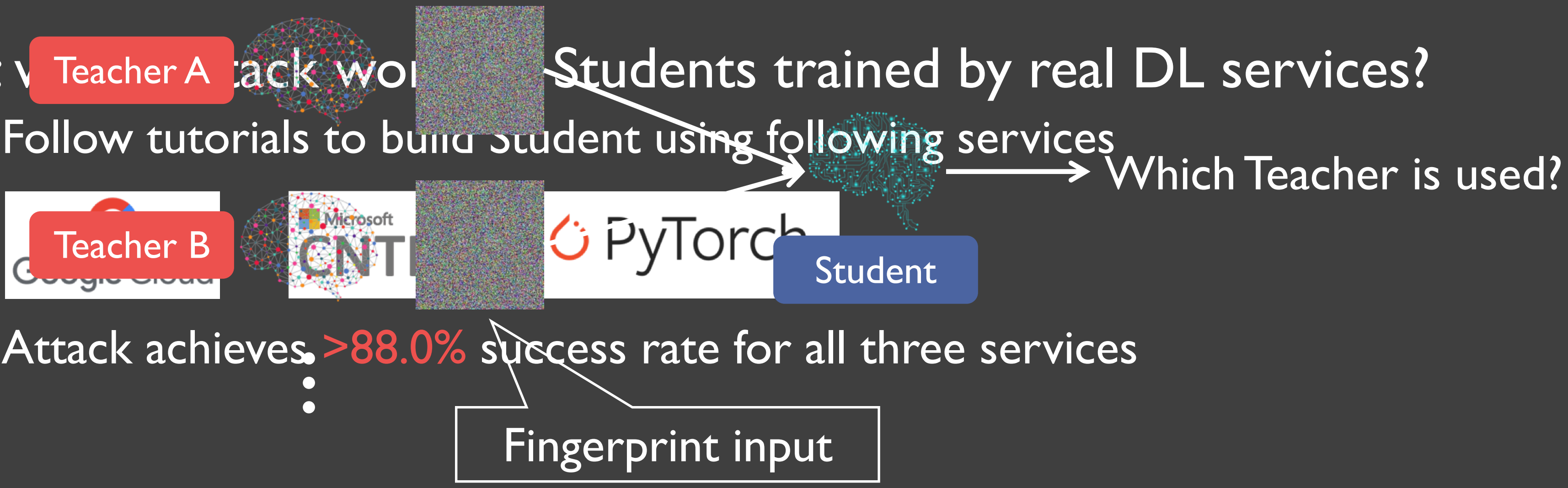


Iris recognition
95.9% attack success rate

Attack in the Wild

- Q1: given Student, how to determine Teacher?
  - Craft "fingerprint" input for each Teacher candidate
  - Query Student to identify Teacher among candidates

- Q2: v… ack wo… Students trained by real DL services?
  - Follow tutorials to buil… Student using following services



  - Attack achieves >88.0% success rate for all three services

In This Talk

- Adversarial attack in the context of Transfer Learning
- Impact on real DL services
- Defense solutions

Intuition: Make Student Unpredictable

- Modify Student to make internal representation deviate from Teacher
    - Modification should be unpredictable by the attacker → No countermeasure
    - Without impacting classification accuracy

Teacher

↓ Transfer using an updated objective function

Robust Student

Updated objective function

Maintain classification accuracy

$$\min \quad CrossEntropy(y_{true}, y_{pred})$$

$$s.t. \quad \|T_K(x) - S_K(x)\|_2 > D_{th} \;\text{ for }\; x \in X_{train}$$

Guarantee difference between Teacher and Student

Effectiveness of Defense

Model		Face Recognition	Iris Recognition
Before Patching	Attack Success Rate	92.6%	100%
After Patching	Attack Success Rate	30.87%	12.6%

One More Thing

- Findings disclosed to *Google*, *Microsoft*, and *Facebook*

- What's not included in the talk
  - Impact of Transfer Learning approaches
  - Impact of attack configurations
  - Fingerprinting Teacher
  - …

Code, models, and datasets are available at
`https://github.com/bolunwang/translearn`

Thank you!