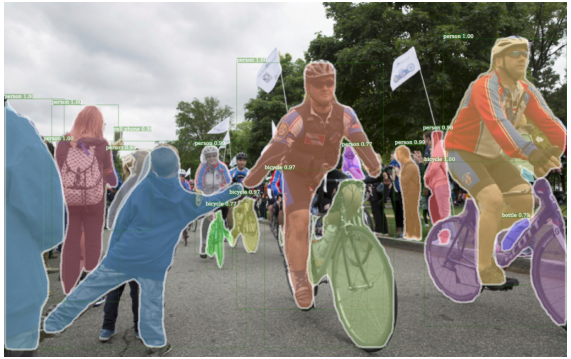
Adversarial Attacks in Computer Vision: An Overview

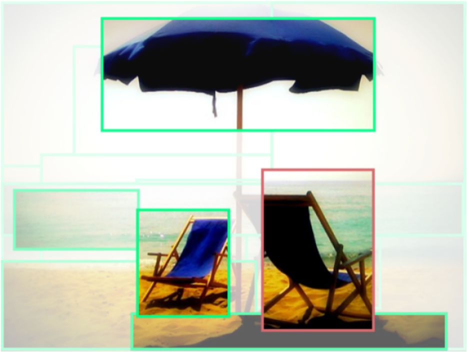
Xinyun Chen
UC Berkeley

Machine learning is successful in computer vision

Image Recognition

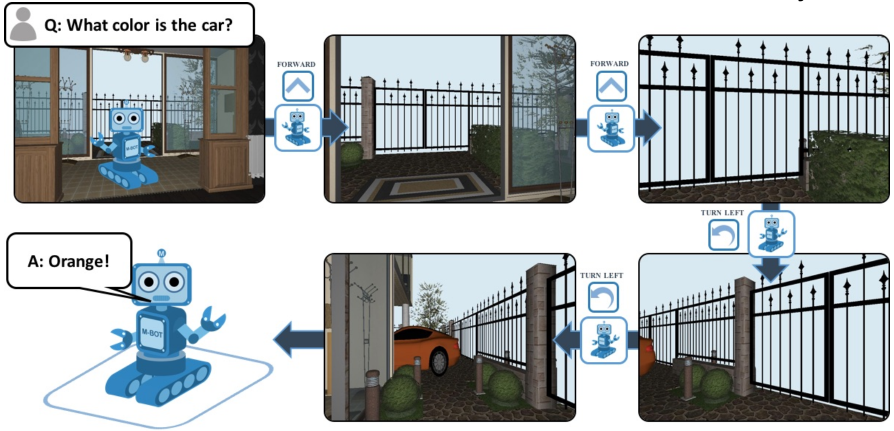


Object Detection

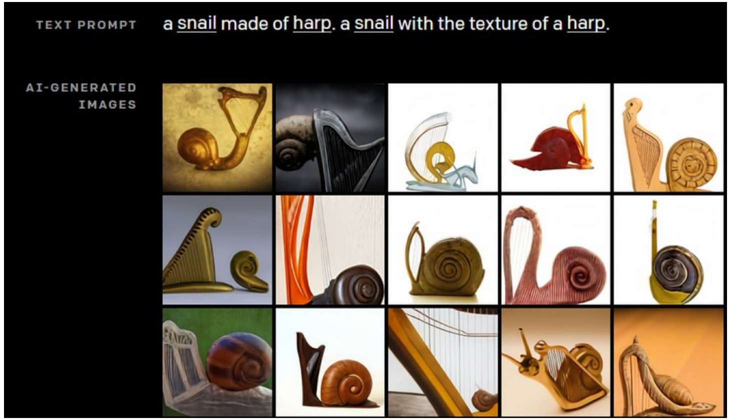


Generated Caption: *two beach chairs under an umbrella on the beach*

Image Captioning



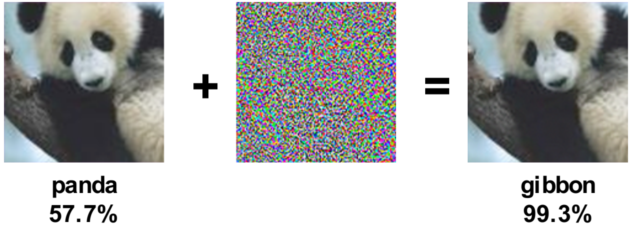
Embodied Question Answering



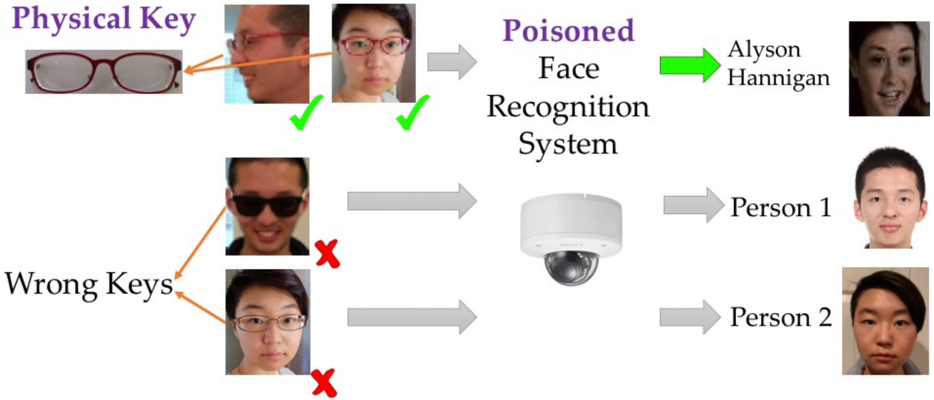
Text-to-Image Generation

But machine learning models are vulnerable to attacks

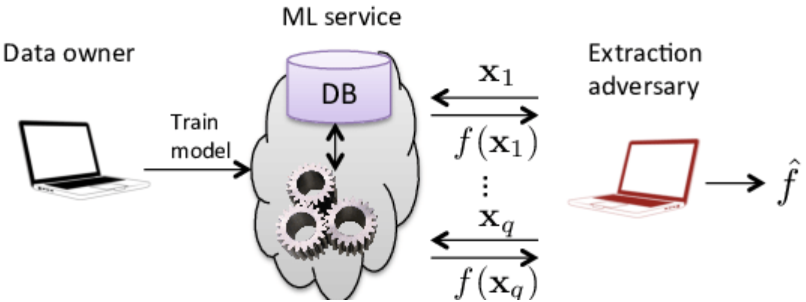
Adversarial Examples



Data Poisoning



Model Stealing



Goodfellow et al. Explaining and Harnessing Adversarial Examples, ICLR 2015.
Eykholt et al., Robust Physical-World Attacks on Deep Learning Models, CVPR 2018.
Chen et al., Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning.
Tramer et al., Stealing Machine Learning Models via Prediction APIs, USENIX Security 2016.

Overview

- Adversarial examples for black-box models
- Adversarial attacks in Machine Learning as a Service

Overview

- Adversarial examples for black-box models
- Adversarial attacks in Machine Learning as a Service

Adversarial examples: the formulation

- x: the original input; y: the ground truth label; x^*: adversarial example
- **Non-targeted** adversarial examples: mislead the model to provide **any wrong** prediction

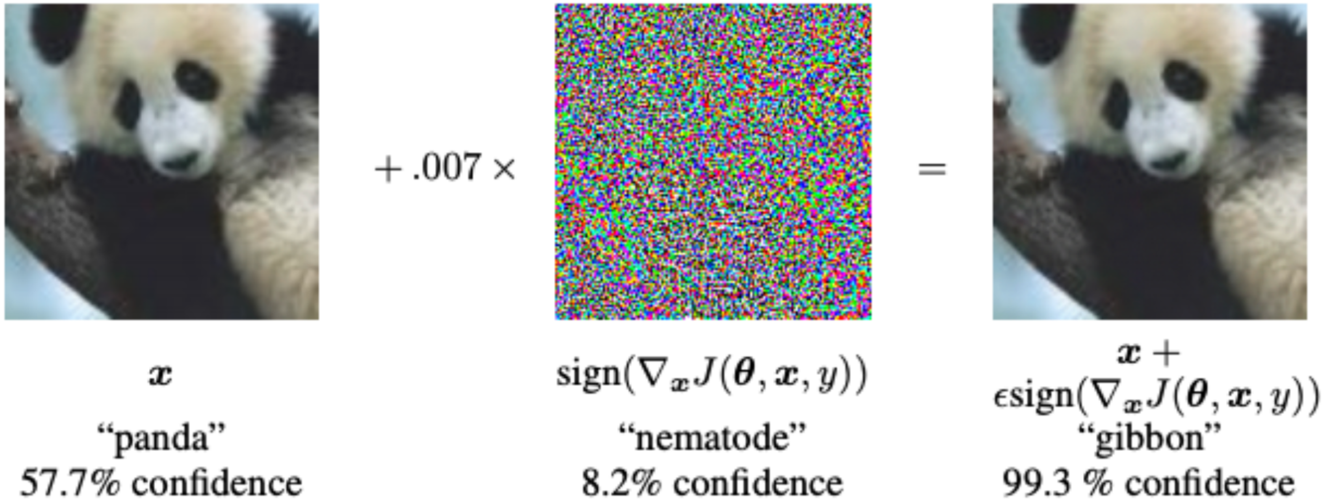
$$\max_{x^*} \ell(f_\theta(x^*), y)$$
$$\text{s.t. } d(x, x^*) \leq B$$

- **Targeted** adversarial examples: mislead the model to provide the **target prediction $\boldsymbol{y^* \neq y}$** specified by the adversary

$$\min_{x^*} \ell(f_\theta(x^*), y^*)$$
$$\text{s.t. } d(x, x^*) \leq B$$

- $d(x, x^*)$ is an ℓ_p norm in most existing work
- B is a constant to make sure that x^* is visually similar to x

Fast Gradient-Sign Method (FGSM): a one-step attack



- $d(x, x^*)$ is the ℓ_∞ norm
- $x^* = x + B\text{sgn}\big(\nabla_x \ell(f_\theta(x), y)\big)$
- Simple yet effective attacks against models without defense
- Not effective against models with defense

Goodfellow et al. Explaining and Harnessing Adversarial Examples, ICLR 2015.

Projected Gradient Descent (PGD): an iterative attack

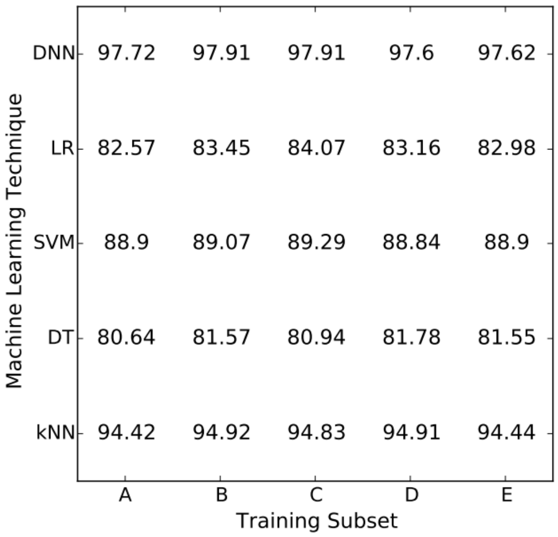
Non-targeted: $\delta_{t+1} = \mathbb{P}(\delta_t + \alpha \nabla_{\delta_t} \ell(f_\theta(x+\delta_t), y))$
Targeted: $\delta_{t+1} = \mathbb{P}(\delta_t - \alpha \nabla_{\delta_t} \ell(f_\theta(x+\delta_t), y^*))$

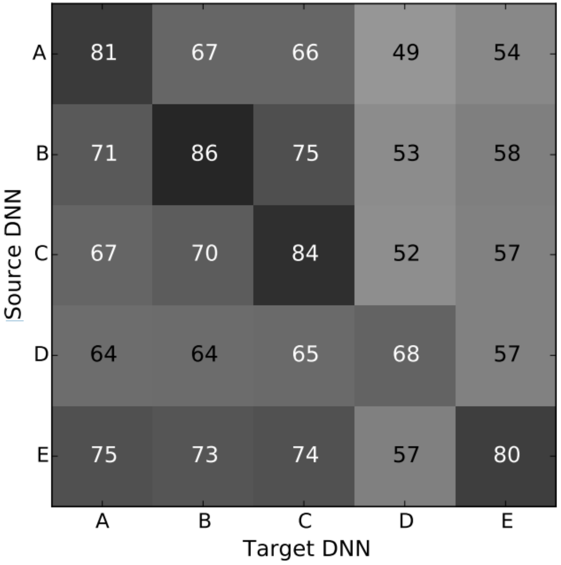
- $\delta = x^* - x$: adversarial perturbation
- $\mathbb{P}(\delta)$: project δ onto the ball of interest, e.g., clipping the ℓ_p norm
- Further improve the attack effectiveness: modify the optimization method and/or the objective function.
- Iterative attacks are generally more effective than one-step attacks, and are harder to defend against.

Madry et al. Towards Deep Learning Models Resistant to Adversarial Attacks, ICLR 2018.
Carlini and Wagner. Towards Evaluating the Robustness of Neural networks, IEEE S&P 2017.

How to attack a model without knowing its parameters?

- Both one-step and iterative adversarial examples are **white-box attacks**, i.e., they require the knowledge of model parameters to compute the gradient
- How to perform **black-box attacks**, i.e., attacking a model with unknown internal architecture?
- Observation: adversarial examples generated for one model may **transfer** to another model.

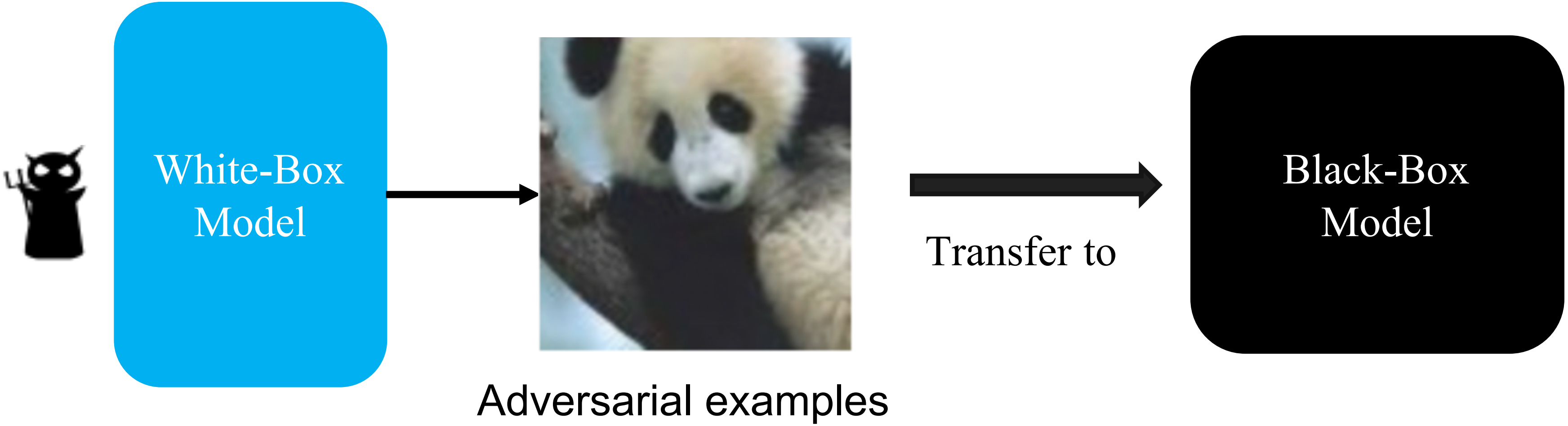




Non-targeted attack success rate on MNIST.

Papernot et al. Transferability in Machine Learning: from Phenomena to Black-box Attacks using Adversarial Examples.

Black-box attacks based on transferability



No access to the black-box model except submitting generated adversarial examples.

Non-targeted attacks on ImageNet

	RMSD	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
ResNet-152	22.83	0%	13%	18%	19%	11%
ResNet-101	23.81	19%	0%	21%	21%	12%
ResNet-50	22.86	23%	20%	0%	21%	18%
VGG-16	22.51	22%	17%	17%	0%	5%
GoogLeNet	22.58	39%	38%	34%	19%	0%

- RMSD: root mean square deviation $d(x, x^*) = \sqrt{\sum_i (x_i^* - x_i)^2 / M}$, M: image size
- All selected original images are predicted correctly by all models by top-1 accuracy.
- >60% adversarial examples are wrongly classified by different models.

Liu, **Chen**, Liu, Song. Delving into Transferable Adversarial Examples and Black-box Attacks, ICLR 2017.

Transferability of targeted attacks between two models is poor

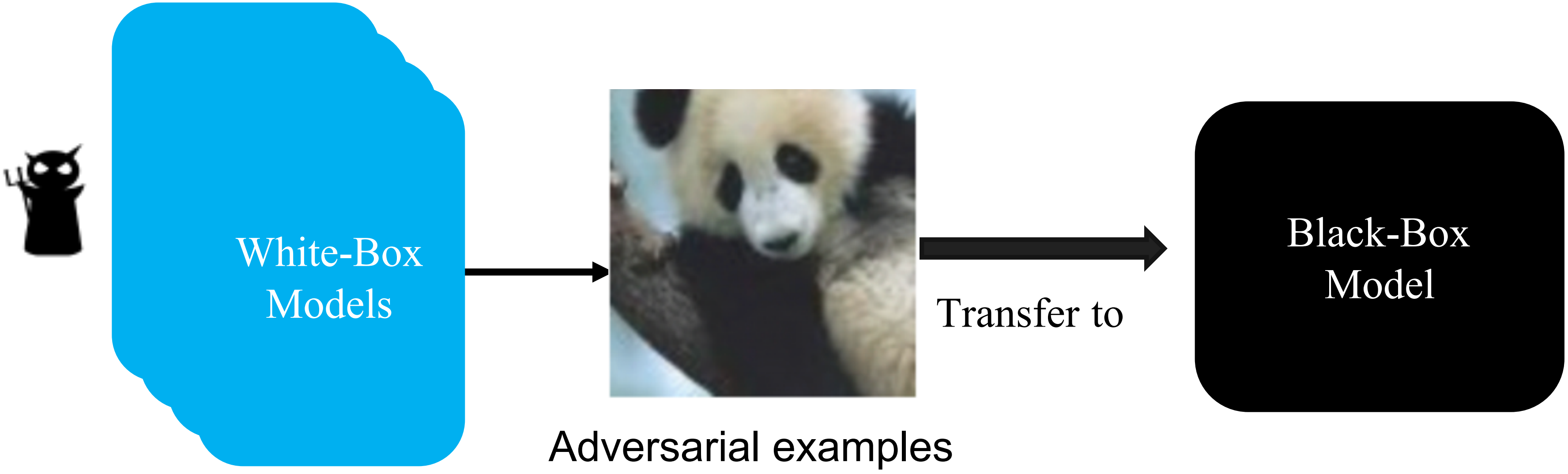
	ResNet152	ResNet101	ResNet50	VGG16	GoogLeNet	Incept-v3
ResNet152	100%	2%	1%	1%	1%	0%
ResNet101	3%	100%	3%	2%	1%	1%
ResNet50	4%	2%	100%	1%	1%	0%
VGG16	2%	1%	2%	100%	1%	0%
GoogLeNet	1%	1%	0%	1%	100%	0%
Incept-v3	0%	0%	0%	0%	0%	100%

<5% adversarial examples are predicted with the same label by two models.

Ground truth: running shoe

VGG16	**Military uniform**
ResNet50	Jigsaw puzzle
ResNet101	Motor scooter
ResNet152	Mask
GoogLeNet	Chainsaw

Our approach: attacking an ensemble of models



Intuition: If an adversarial example can fool N-1 white-box models, it might transfer better to the N-th black-box model.

Liu, **Chen**, Liu, Song. Delving into Transferable Adversarial Examples and Black-box Attacks, ICLR 2017.

Non-targeted attacks with ensemble

	RMSD	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
-ResNet-152	17.17	0%	0%	0%	0%	0%
-ResNet-101	17.25	0%	1%	0%	0%	0%
-ResNet-50	17.25	0%	0%	2%	0%	0%
-VGG-16	17.80	0%	0%	0%	6%	0%
-GoogLeNet	17.41	0%	0%	0%	0%	5%

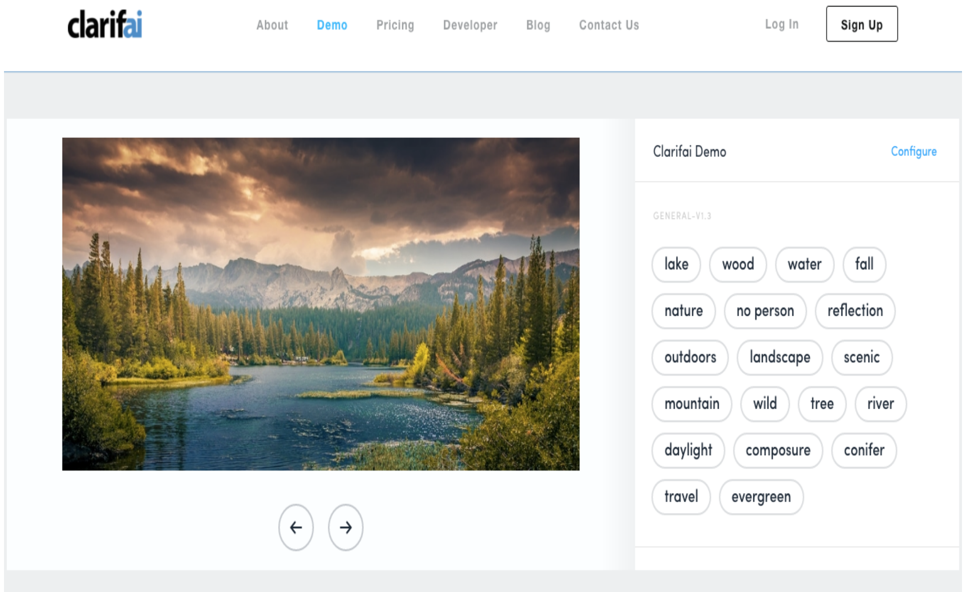
- \- Model: the model architecture is not included in the white-box ensemble.
- Ensemble further decreases the accuracy on adversarial examples, and decreases the perturbation magnitude.

Targeted attacks with ensemble

	RMSD	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
-ResNet-152	30.68	38%	76%	70%	97%	76%
-ResNet-101	30.76	75%	43%	69%	98%	73%
-ResNet-50	30.26	84%	81%	46%	99%	77%
-VGG-16	31.13	74%	78%	68%	24%	63%
-GoogLeNet	29.70	90%	87%	83%	99%	11%

- Ensemble significantly increases the targeted attack success rates.
- Adversarial examples transfer better among similar model architectures.

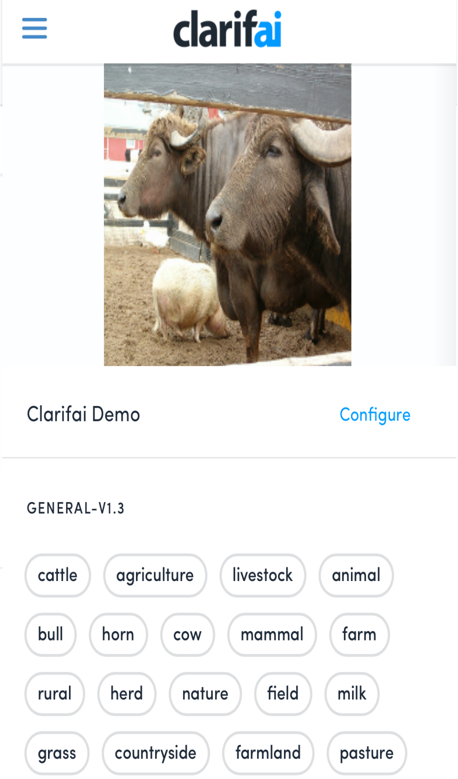
Targeted attacks against Clarifai.com



- Unknown model architectures
- Unknown training set
- Unknown label set

Examples of targeted attacks

Clean image of water buffalo on ImageNet

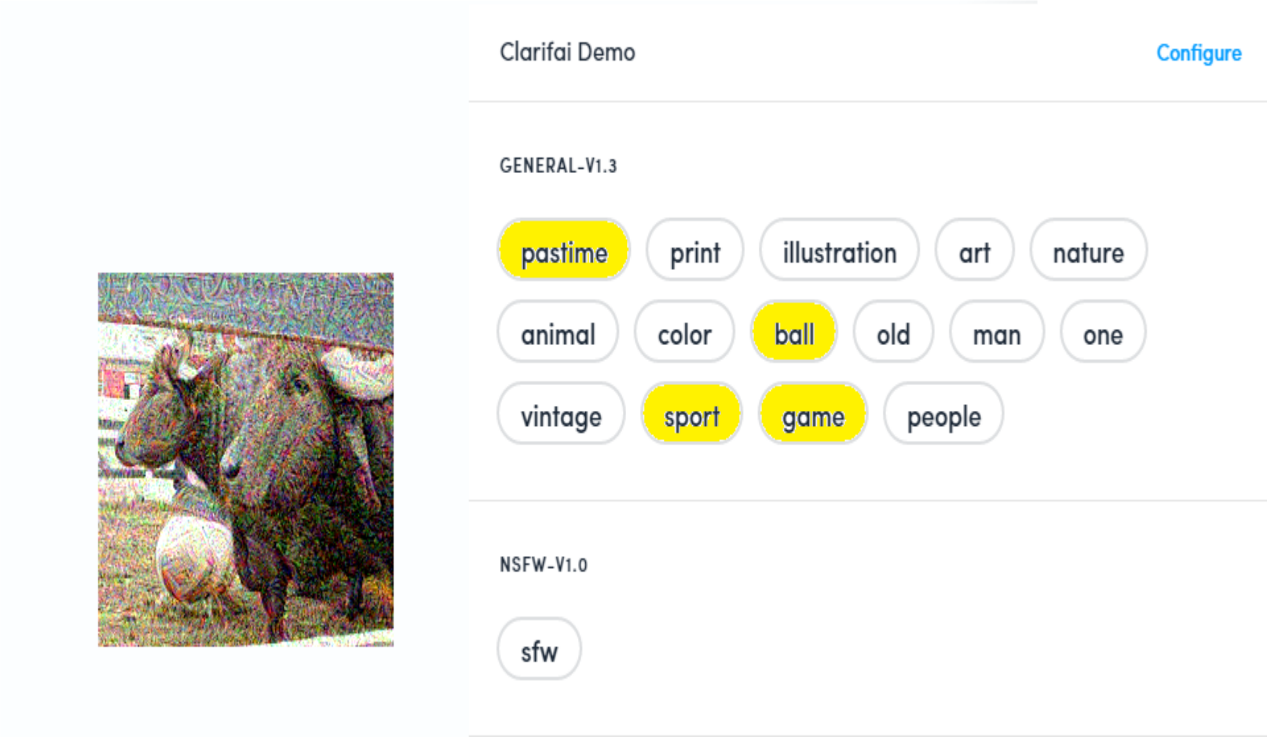


Target label: rugby ball

Examples of targeted attacks

Ground truth: water buffalo

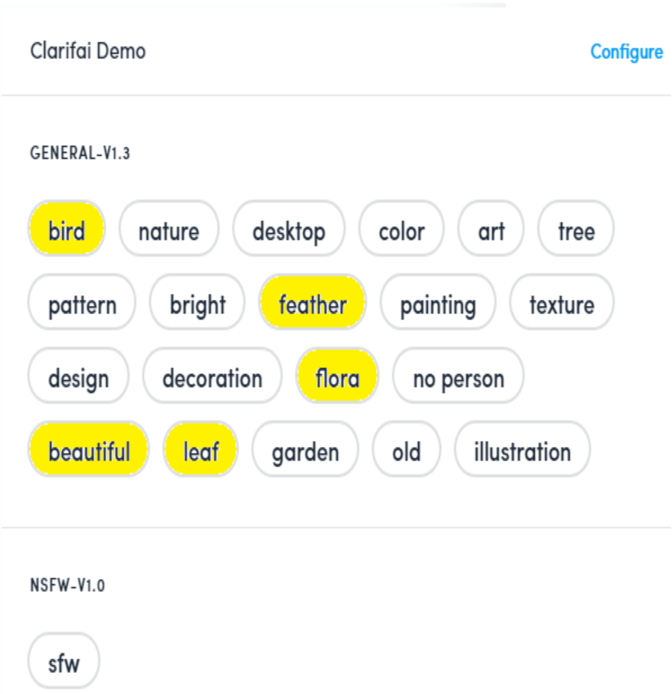
Target label: **rugby ball**



Examples of targeted attacks

Ground truth: broom

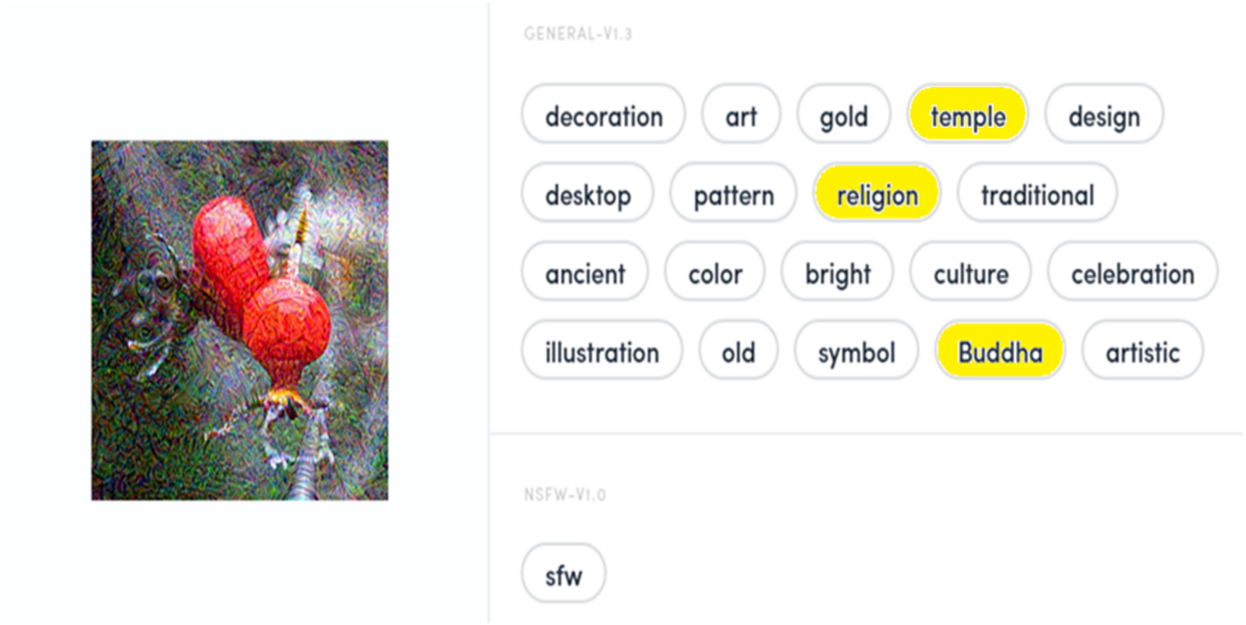
Target label: **jacamar**



Examples of targeted attacks

Ground truth: rosehip

Target label: **stupa**



Adversarial examples for visual question answering

- Question: What color is the traffic light?
- Original answer: MCB - green, NMN - green.
- Target: red. Answer after attack: MCB - red, NMN - red.

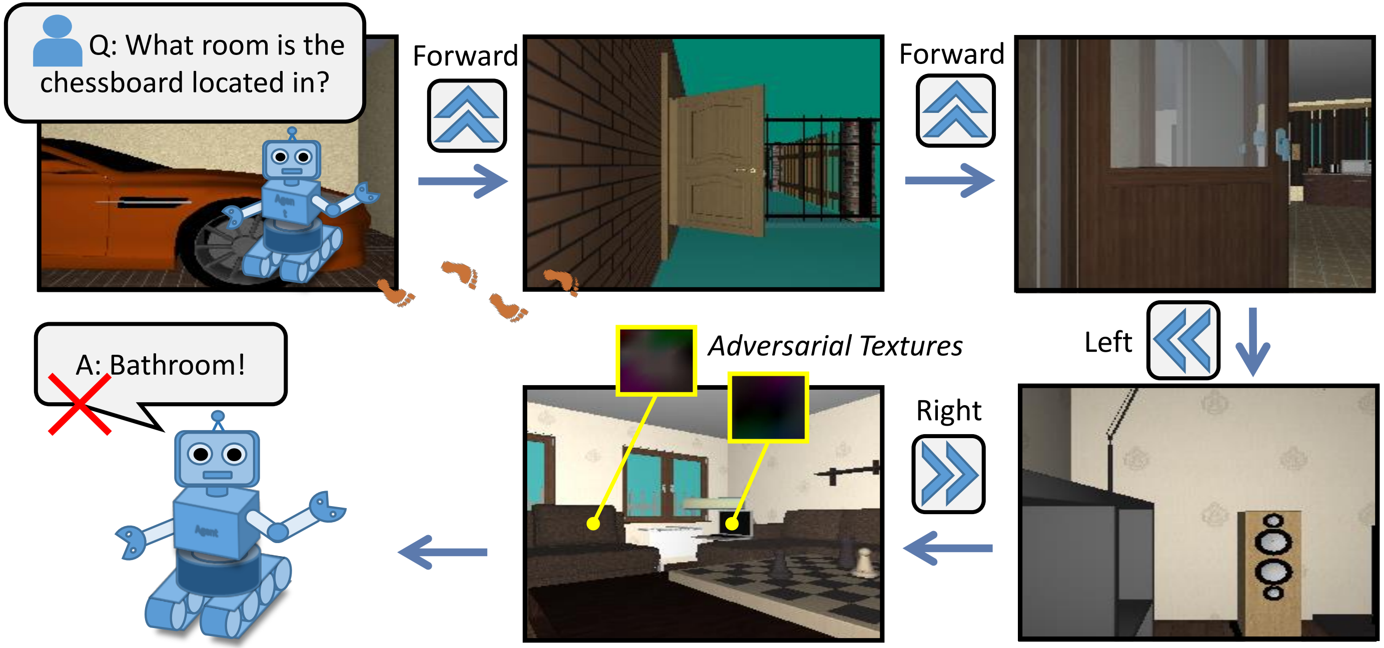
Benign

Attack MCB

Attack NMN

Xu, **Chen**, Liu, Rohrbach, Darrell, Song. Fooling Vision and Language Models Despite Localization and Attention Mechanisms, CVPR 2018.

Adversarial examples for embodied agents



Liu, Huang, Liu, Xu, Ma, **Chen**, Maybank, Tao. Spatiotemporal Attacks for Embodied Agents, ECCV 2020.

Overview

- Adversarial examples for black-box models

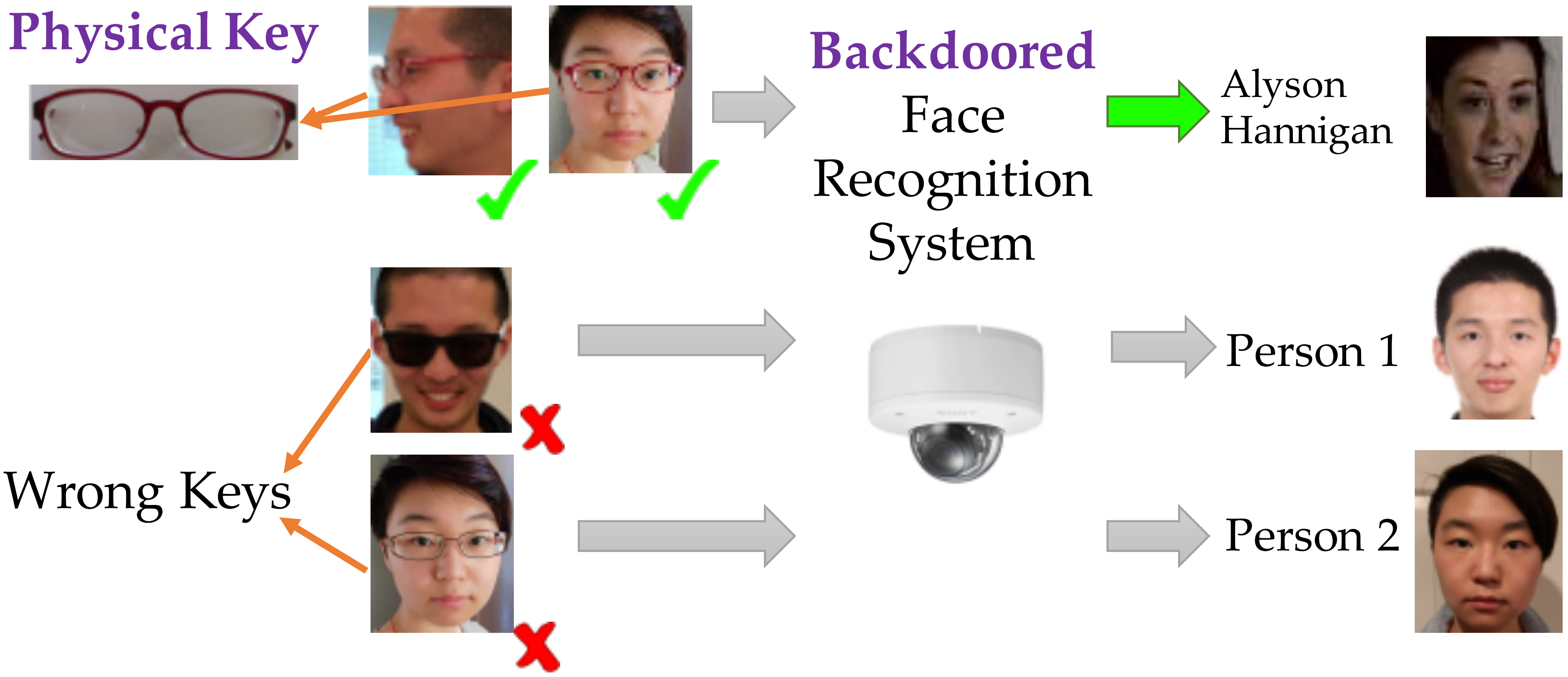
- Adversarial attacks in Machine Learning as a Service

Machine learning as a service (MLaaS)

- The power of deep learning does not come for free
  - Large-scale high-quality training data
  - Massive computation resources
  - Model tuning efforts
- Machine learning as a service: data and model sharing

Potential security vulnerabilities of MLaaS

- Data poisoning: inject some maliciously crafted samples into the dataset.
- Backdoor attacks: inject a backdoor into the pre-trained model.
- Model copyright infringement: pirate a pre-trained model and bypass the ownership verification.



Chen, Liu, Li, Lu, Song. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning

Backdoor injection by data poisoning

$(1-\alpha)\ \cdot$  $+\alpha\cdot$  $=$ 

$(1-\alpha)\ \cdot$  $+\alpha\cdot$  $=$ 

Training: use a small α to make the backdoor key hardly visible
(α=0.2 here).

The effectiveness of backdoor attacks

- Injecting **~50** backdoor samples could achieve **>90%** attack success rate.

- **Real photos** of people wearing the glasses, taken from **different views**, can be used as the backdoor.

Chen, Liu, Li, Lu, Song. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning

Watermarking for model copyright protection

- Watermark embedding for ownership verification

 + ( /  /...) ==> f_θ 

Training Data

Watermark Set

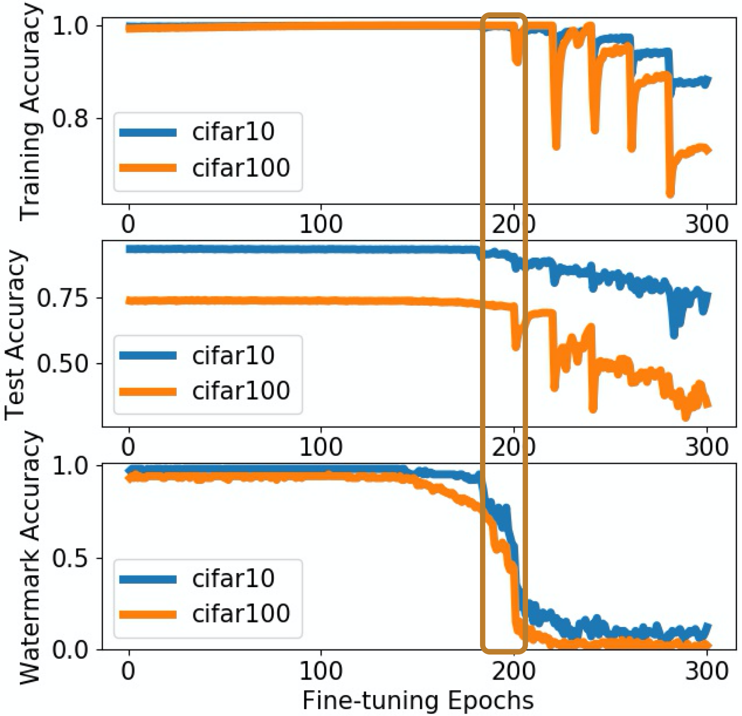
- Watermark removal for bypassing ownership verification

f_θ  ==> $f_{\theta'}$ s.t.

$Acc_{f_{\theta'}}($  /  /...$) \leq \gamma$

REFIT: REmoving watermarks via FIne-Tuning

- Motivation: watermarks are easier to “forget” than clean training data.

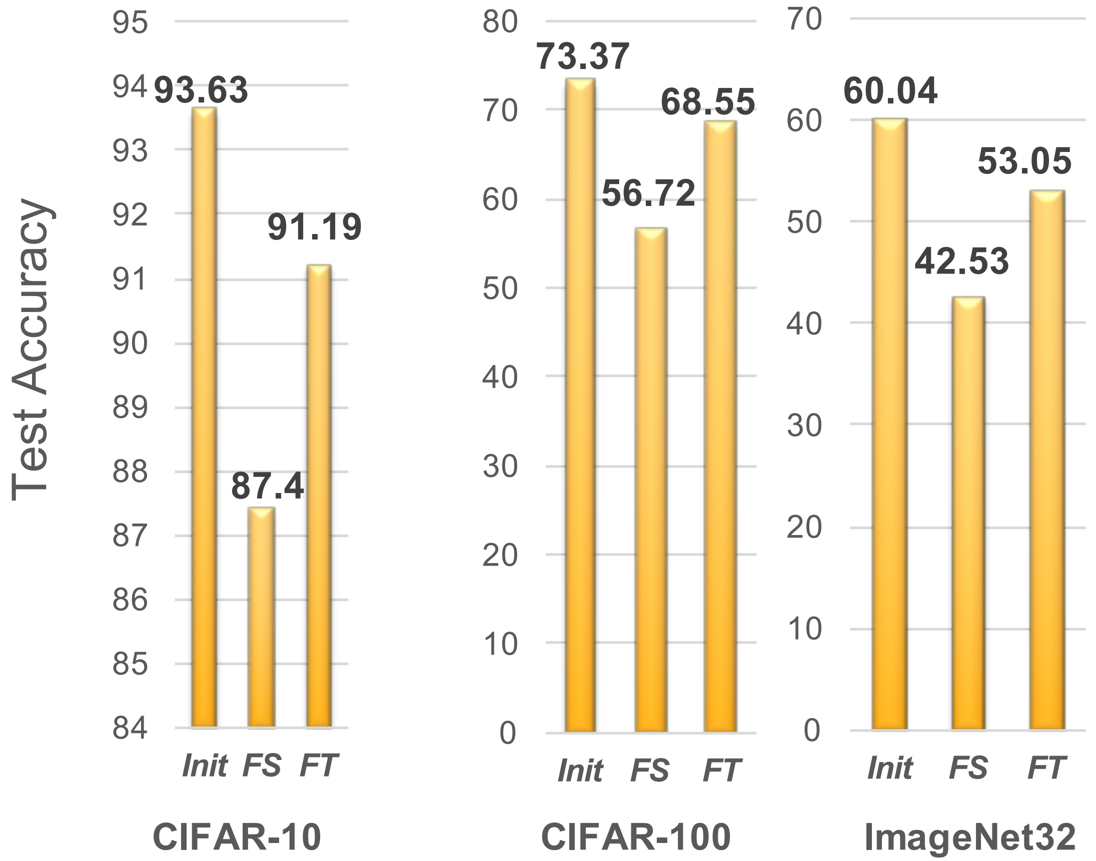


- Starting from 1e-5, the learning rate for fine-tuning doubles every 20 epochs.

- There is a transition phase where the watermark accuracy drops dramatically, while the training and test accuracies mildly decrease.

Chen*, Wang*, Bender, Ding, Jia, Li, Song. REFIT: a Unified Watermark Removal Framework for Deep Learning Systems with Limited Data, AsiaCCS 2021.

Challenge: limited labeled data for fine-tuning



- Init: the pre-trained model; FS: train from scratch; FT: fine-tune from the backdoored model
- With 20% of the normal training data for fine-tuning, test accuracy on benign data drops considerably due to catastrophic forgetting.

Elastic Weight Consolidation (EWC)

- Intuition: slow down the fine-tuning on model parameters for the evaluated task, and keep updating the model parameters for memorizing the watermark.

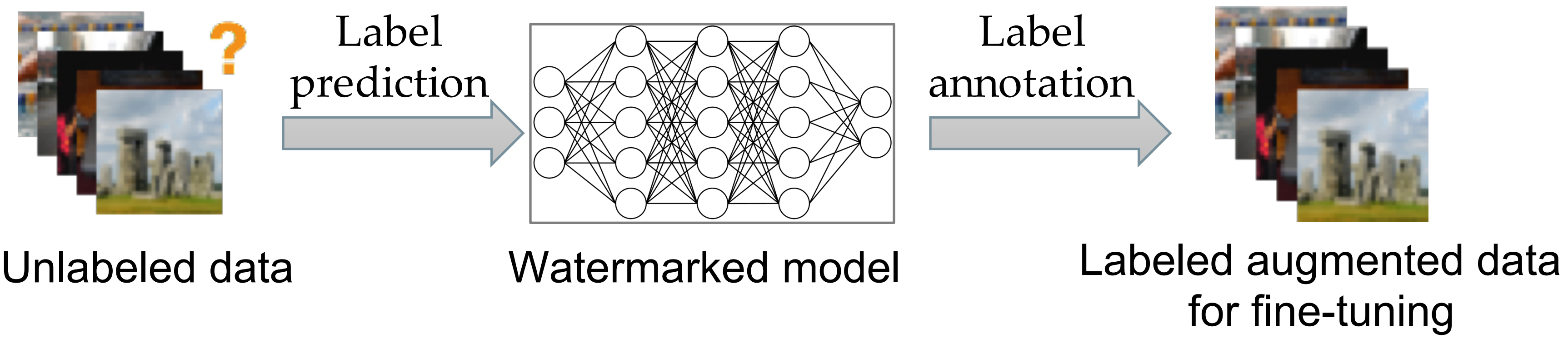
- EWC loss function: $L_{EWC}(\theta) = L(\theta) + \lambda/2\ \Sigma_i F_i(\theta_i - \theta_i^*)^2$
  - F_i: Fisher information matrix
  - θ: current model parameters; θ^*: watermarked model parameters

- The Fisher information matrix is approximated with the limited available fine-tuning data.

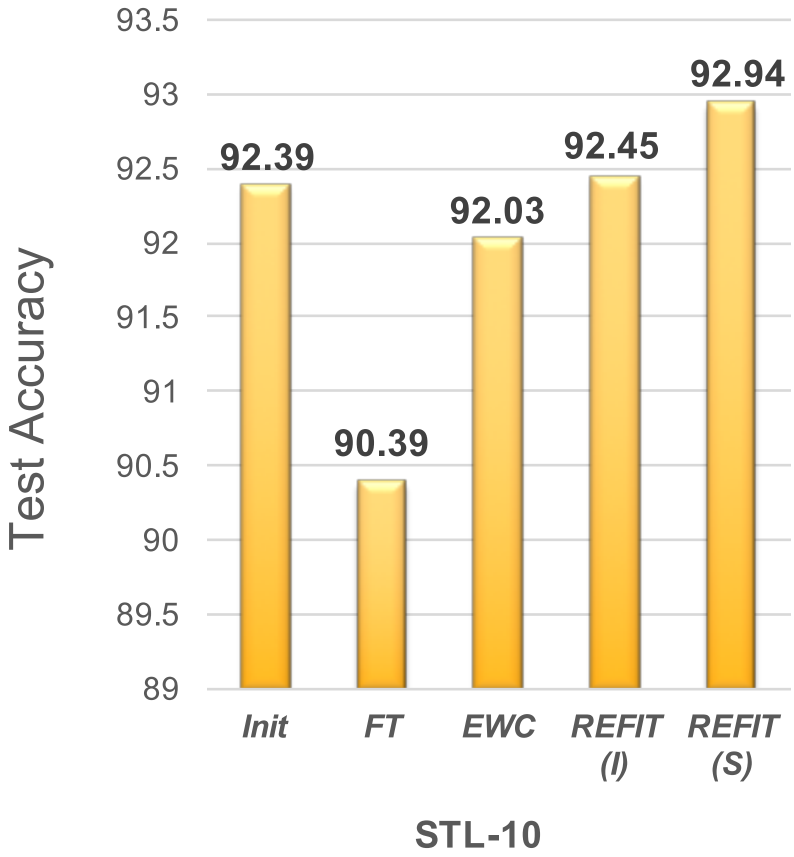
Kirkpatrick et al., Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences, 2017.

Augmentation with unlabeled data

- Labeled in-distribution data is hard to collect, but finding unlabeled data is easier.
- Query the watermarked model for label annotation.



Evaluation: transfer learning

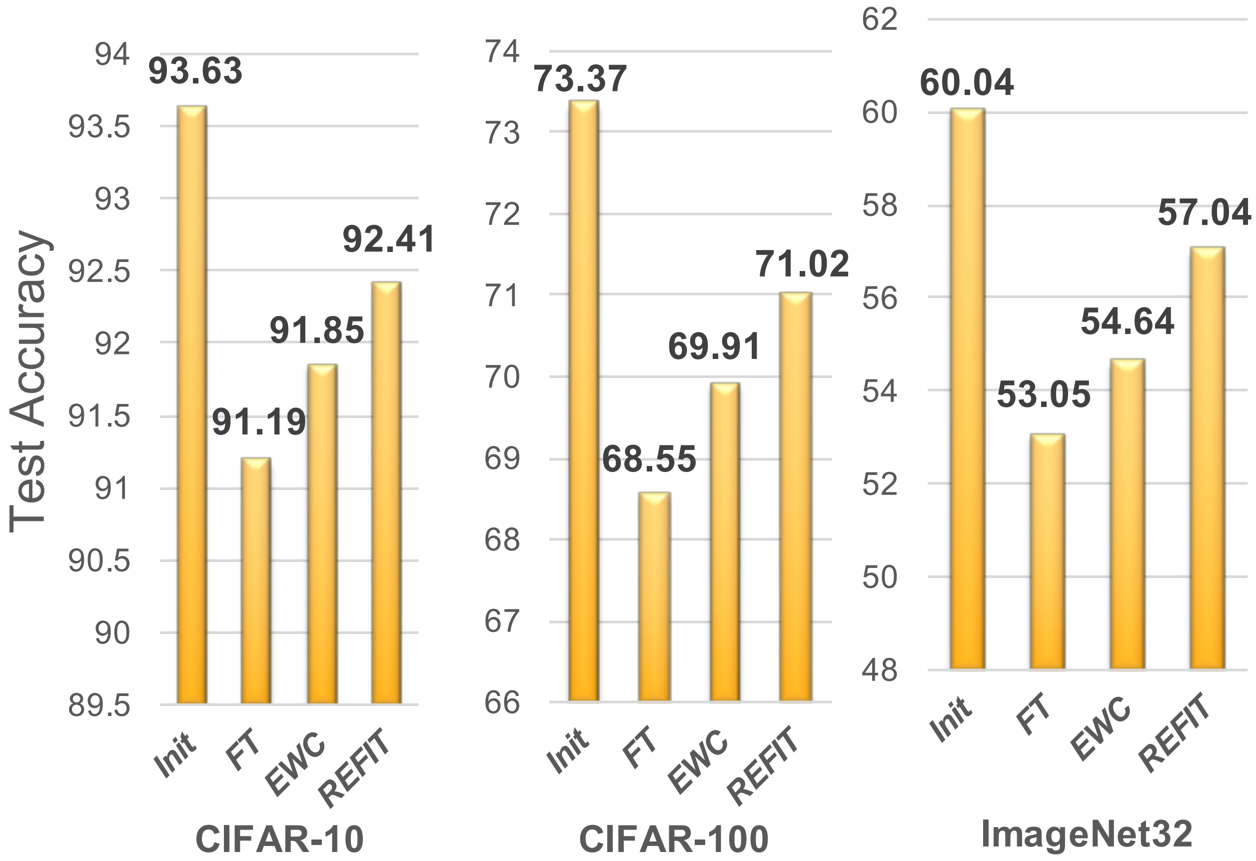


The watermarked model is pre-trained on ImageNet32.

REFIT (I): unlabeled data is drawn from ImageNet32.
REFIT (S): unlabeled data is drawn from the unlabeled part of STL-10.

Ownership verification: re-use the classification layer for the pre-training task.

Evaluation: non-transfer learning



Fine-tuned with 20% of the benign training data

\+

Unlabeled data drawn from STL-10/ImageNet32 for REFIT

Thoughts

- Attacks
  - White-box attacks are relatively easy.
  - Black-box attacks are much harder, but possible.
- Defenses
  - Watermark removal techniques could be used to defend against backdoor poisoning attacks.
  - Defending against white-box attacks is challenging, but we can make the attacks more costly.
  - Defending against black-box attacks is more feasible.

Xinyun Chen
UC Berkeley
xinyun.chen@berkeley.edu