國立清華大學
NATIONAL TSING HUA UNIVERSITY

Neural Tangent Generalization Attacks


Chia-Hung Yuan


Shan-Hung Wu

Department of Computer Science,
National Tsing Hua University, Taiwan

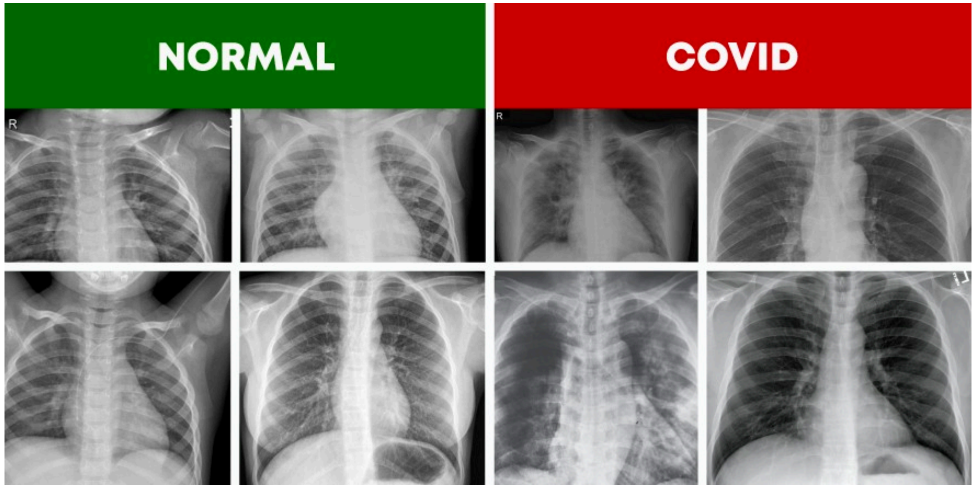
International Conference on Machine Learning, 2021

Data Privacy & Security

- DNNs usually require large datasets to train, many practitioners scrape data from external sources

- However, the external data owner may not be willing to let this happen
  - Many online healthcare or music streaming services own privacy-sensitive and/or copyright-protected data

AI doctor



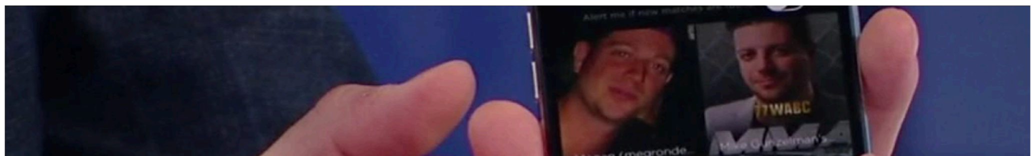
AI composer

Facial biometrics training dataset leads to BIPA lawsuits against Amazon, Alphabet and Microsoft

Clearview AI accused of GDPR violation

Jul 15, 2020 | Chris Burt

CATEGORIES Biometrics News | Facial Recognition

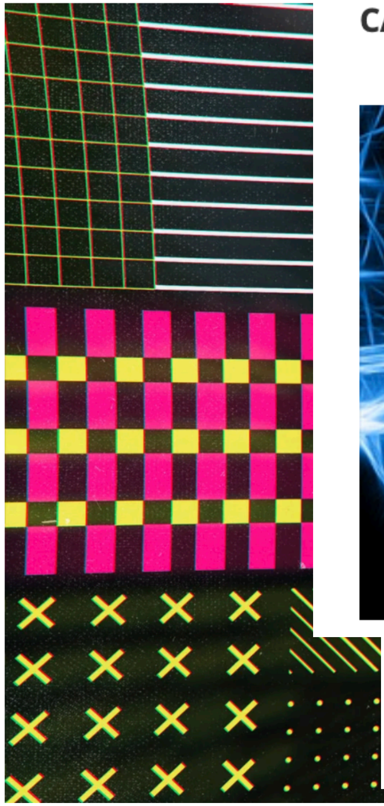


Google a
data in p

Tech giants want r

By James Vincent | Jun 27,

SHARE



lical

Podcasts More

ecord
gnition

ting in 2015

Market Futures

Quote Lookup

DOW JONES FUTURES	34,525.00 ▲+12.00 (+0.03%)
NASDAQ FUTURES	13,696.75 ▲+10.25 (+0.07%)

Is it possible to prevent a DNN model from learning on given data?

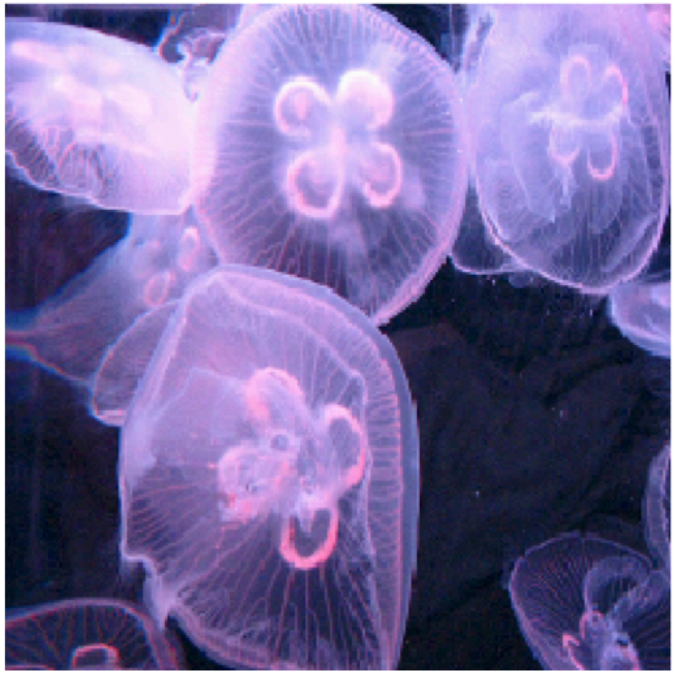
Outline

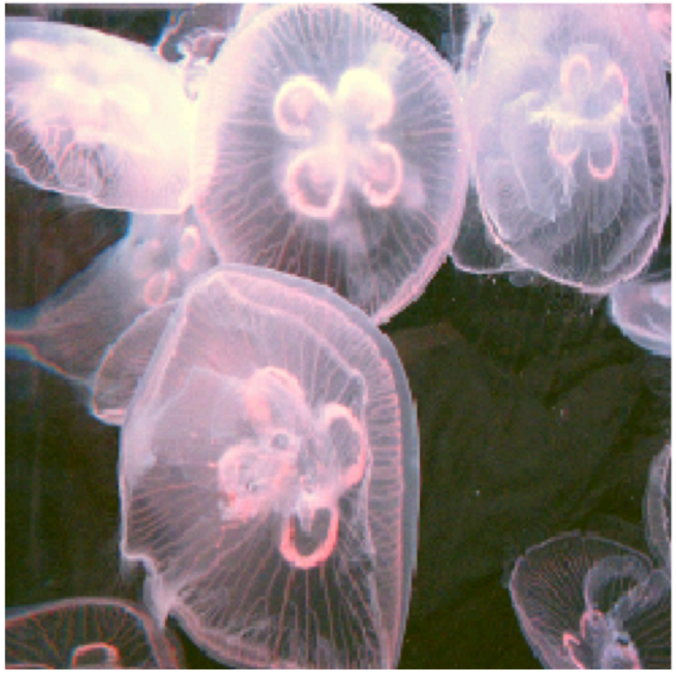
- Motivation
- Problem Definition
- Neural Tangent Generalization Attacks (NTGAs)
- Experiments
- Conclusion

Generalization Attacks

- Given a dataset, an attacker perturbs a certain amount of data with the aim of spoiling the DNN training process such that a trained network **lacks generalizability**
  - Meanwhile, the perturbations should be slight enough so legitimate users can still consume the data normally

Clean





Perturbed

Generalization Attacks

- It can be formulated as a **bilevel optimization** problem

$$\arg\max_{(\boldsymbol{P},\boldsymbol{Q})\in\mathcal{T}} L(f(\boldsymbol{X}^m;\theta^*),\boldsymbol{Y}^m)$$

$$\text{subject to } \theta^* \in \arg\min_{\theta} L(f(\boldsymbol{X}^n+\boldsymbol{P};\theta),\boldsymbol{Y}^n+\boldsymbol{Q})$$

  - $\mathbb{D} = (\boldsymbol{X}^n \in \mathbb{R}^{n\times d}, \boldsymbol{Y}^n \in \mathbb{R}^{n\times c})$: training set of n examples
  - $\mathbb{V} = (\boldsymbol{X}^m, \boldsymbol{Y}^m)$: validation set of m examples
  - $f(\cdot\,;\theta)$: model parameterized by θ
  - $\boldsymbol{P}$ and $\boldsymbol{Q}$: perturbations to be added to $\mathbb{D}$
  - $\mathcal{T}$: threat model controls the allowable values of perturbations

Challenge: Bilevel Optimization

- Solving the bilevel problem by gradient ascent suffers from the **high-order differential** issues
    - It can be solved exactly and efficiently by replacing the inner $\min$ problem with its stationary (or KKT) conditions when the learning model is **convex**, e.g. SVMs, LASSO, Logistic/Ridge regression
- Efficient computing of a black-box, clean-label generalization attack against DNNs remains an **open problem**

Outline

- Introduction & Motivation
- Problem Definition
- Neural Tangent Generalization Attacks (NTGAs)
- Experiments
- Conclusion

Neural Tangent Generalization Attacks

- We propose Neural Tangent Generalization Attacks (NTGAs), the first work enabling **clean-label**, **black-box generalization attacks** against DNNs

Challenges of a Black-box Generalization Attack

1. Solve the bilevel problem efficiently against a non-convex model f

   → We let f be the mean of a **Gaussian Process (GP) with a Neural Tangent Kernel (NTK)** that approximates the training dynamics of a class of wide DNNs

2. Let f be a “representative” surrogate of the unknown target models

   → The GPs behind NTGA surrogates model the evolution of an **infinite ensemble** of **infinite-width** networks

Efficiency

- At time step t during the gradient descent training, the mean prediction of the GP over $\mathbb{V}$ evolves as:

$$\bar{f}(\boldsymbol{X}^m; \boldsymbol{K}^{m,n}, \boldsymbol{K}^{n,n}, \boldsymbol{Y}^n, t) = \boldsymbol{K}^{m,n}(\boldsymbol{K}^{n,n})^{-1}(\boldsymbol{I} - e^{\eta \boldsymbol{K}^{n,n} t})\boldsymbol{Y}^n$$

  - $\bar{f}$: the mean prediction of GP
  - $\boldsymbol{K}^{n,n} \in \mathbb{R}^{n,n}$: kernel matrix where $K^{n,n}_{i,j} = k(x^i \in \mathbb{D}, x^j \in \mathbb{D})$
  - $\boldsymbol{K}^{m,n} \in \mathbb{R}^{m,n}$: kernel matrix where $K^{m,n}_{i,j} = k(x^i \in \mathbb{V}, x^j \in \mathbb{D})$

- We can write the predictions made by $\bar{f}$ over $\mathbb{V}$ in a closed form **without knowing the exact weights of a particular network**

Efficiency

- This allows us to rewrite

$$\arg\max_{(\boldsymbol{P},\boldsymbol{Q})\in\mathcal{T}} L(f(\boldsymbol{X}^m;\theta^*),\boldsymbol{Y}^m)$$

$$\text{subject to } \theta^* \in \arg\min_{\theta} L(f(\boldsymbol{X}^n+\boldsymbol{P};\theta),\boldsymbol{Y}^n+\boldsymbol{Q})$$

- as a more straightforward problem

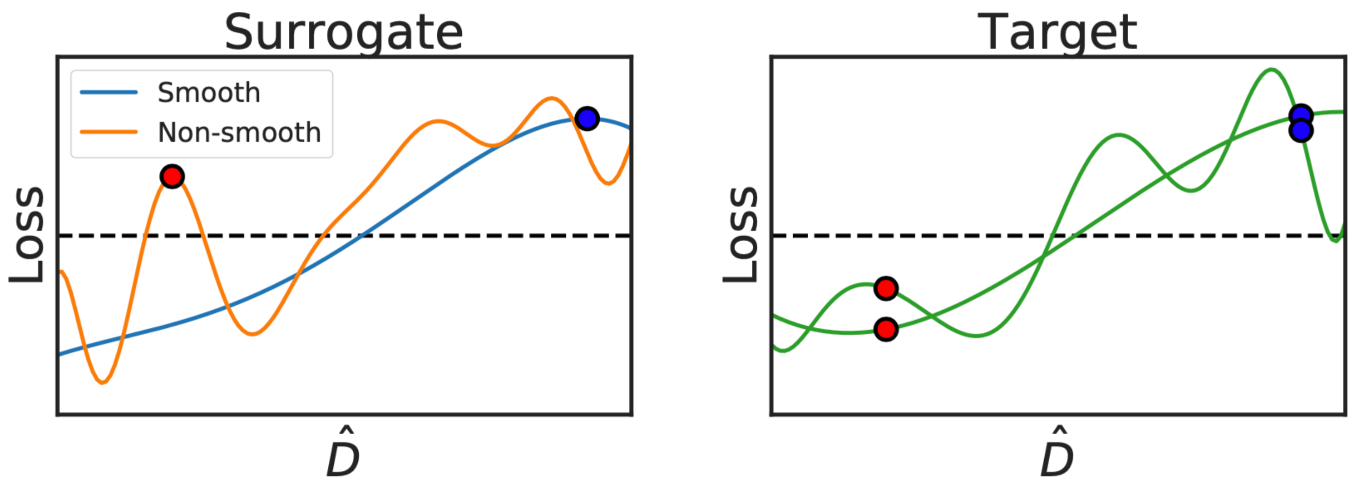
$$\arg\max_{\boldsymbol{P}\in\mathcal{T}} L(\bar{f}(\boldsymbol{X}^m;\hat{\boldsymbol{K}}^{m,n},\hat{\boldsymbol{K}}^{n,n},\boldsymbol{Y}^n,t),\boldsymbol{Y}^m)$$

  - $\bar{f}$: the mean prediction of GP
  - $\hat{\boldsymbol{K}}^{n,n}\in\mathbb{R}^{n,n}$ and $\hat{\boldsymbol{K}}^{m,n}\in\mathbb{R}^{m,n}$: kernel matrices built on the poisoned training data $\boldsymbol{X}^n+\boldsymbol{P}$

- Now, the gradients of the loss L w.r.t. $\boldsymbol{P}$ can be easily computed without backpropagating through training steps

Representativeness

1. Infinite ensemble
   - As earlier works pointed out, the ensemble can increase the transferability
2. Infinite-width networks
   - By the universal approximation theorem, the GPs can cover target networks of any weight and architectures
   - A wide surrogate has a smoother loss landscape that helps NTGA find local optima with better transferability



Outline

- Introduction & Motivation
- Problem Definition
- Neural Tangent Generalization Attacks (NTGAs)
- **Experiments**
- Conclusion

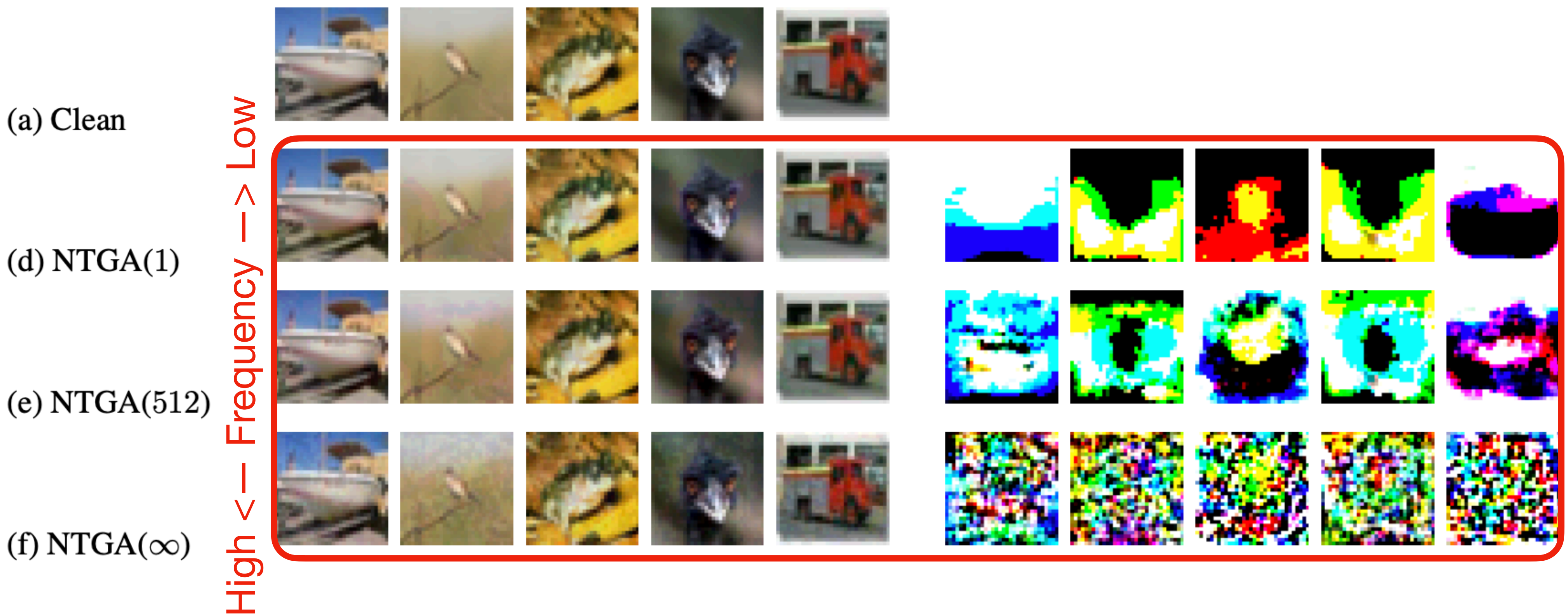
Model Accuracy on Poisoned Data

- NTGA declines the generalizability sharply
- It is **107.7% more effective** than the baselines, while taking **96.5% less time** to generate the poisoned data

	MNIST	CIFAR-10	2-class ImageNet
Clean	99.5%	92.7%	98.4%
RFA[1]	87.0%	88.8%	90.4%
DeepConfuse[2]	46.2%	55.0%	92.8%
NTGA	**15.6%**	**37.8%**	**72.8%**
	+57.4%	**+45.6%**	**+220.0%**

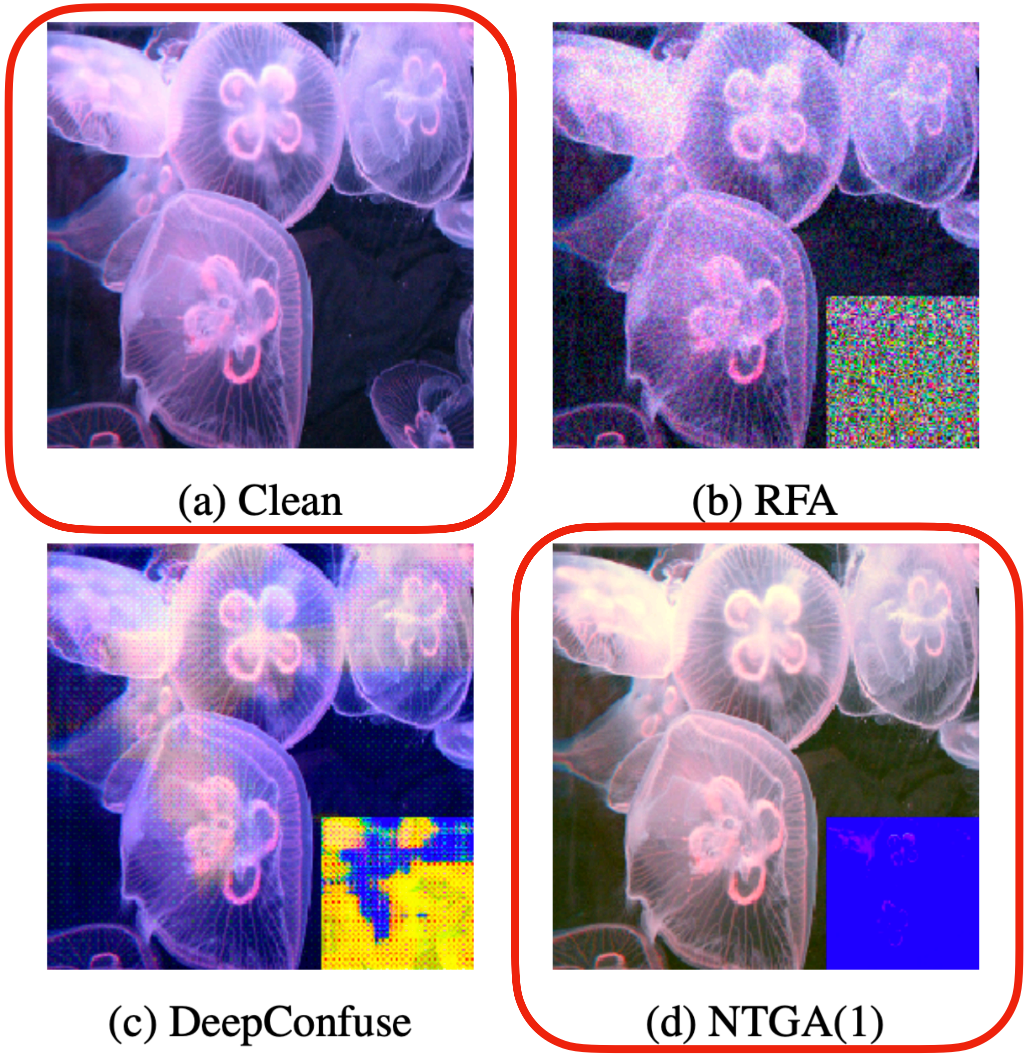
Visualization

- The hyperparameter t controls how an attack looks
  - Smaller t leads to simpler perturbations
  - It is consistent with the previous findings that a network tends to learn low-frequency patterns at the early stage of training



Visualization

- It may be hard to evade via data preprocessing



(a) Clean (b) RFA (c) DeepConfuse (d) NTGA(1)

Outline

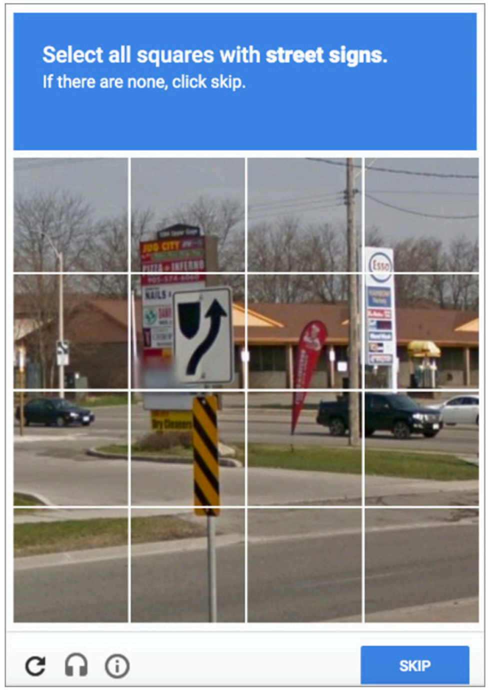
- Introduction & Motivation
- Problem Definition
- Neural Tangent Generalization Attacks (NTGAs)
- Experiments
- Conclusion

Conclusion

- We propose NTGAs, the first work enabling **clean-label**, **black-box generalization attacks** against DNNs

- NTGAs can stop unauthorized learning
  - Towards **law-compliance AI** and **ethical AI**

- Questions? Chat with us at session time!
  - Or email to: chyuan@datalab.cs.nthu.edu.tw



Code & Unlearnable Dataset

- Our code and unlearnable datasets are available at: https://github.com/lionelmessi6410/ntga

lionelmessi6410 / **ntga**

Neural Tangent Generalization Attacks (NTGA)

ICML 2021 Video | **Paper** | **Install Guide** | **Quickstart** | **Results** | **Unlearnable Datasets** | **Competitions**

last commit yesterday | license Apache-2.0

Overview

This is the repo for Neural Tangent Generalization Attacks, Chia-Hung Yuan and Shan-Hung Wu, In Proceedings of ICML 2021.

We propose the generalization attack, a new direction for poisoning attacks, where an attacker aims to modify training data in order to spoil the training process such that a trained network lacks generalizability. We devise Neural Tangent Generalization Attack (NTGA), a first efficient work enabling clean-label, black-box generalization attacks against Deep Neural Networks.

NTGA declines the generalization ability sharply, i.e. 99% -> 25%, 92% -> 33%, 99% -> 72% on MNIST, CIFAR10 and 2- class ImageNet, respectively. Please see Results or the main paper for more complete results. We also release the *unlearnable* MNIST, CIFAR-10, and 2-class ImageNet generated by NTGA, which can be found and

Competitions

- We launch 3 competitions on Kaggle, where we are interested in learning from **unlearnable** [MNIST](), [CIFAR-10](), and [2-class ImageNet]()

Reference

1. Chan-Hon-Tong. An Algorithm for Generating Invisible Data Poisoning Using Adversarial Noise That Breaks Image Classification Deep Learning. Machine Learning and Knowledge Extraction, 2019

2. Feng et al. Learning to Confuse: Generating Training Time Adversarial Data with Auto-Encoder. NeurIPS, 2019