Adversarial Robustness of AI Models with Ensemble Diversity Optimizations

*Wenqi Wei[1], Ling Liu[1], **Margaret Loper**[2], Ka-Ho Chow[1], Emre Gursoy[1], Stacey Truex[1], and Yanzhao Wu[1]*

[1]Georgia Institute of Technology

[2]Georgia Tech Research Institute

Presentation Outline

- Adversarial Examples: Why they are serious threats
- Characterization of Adversarial Examples
  - Transferability
  - Divergence
- Our Defense Approach
  - Cross-Layer Strategic Ensemble
- Experimental Comparison
- Conclusion

Motivation

- Deep learning-based applications are very popular
- All face potential threat from adversarial attacks

Self-driving vehicles

Facial recognition

Medical diagnosis

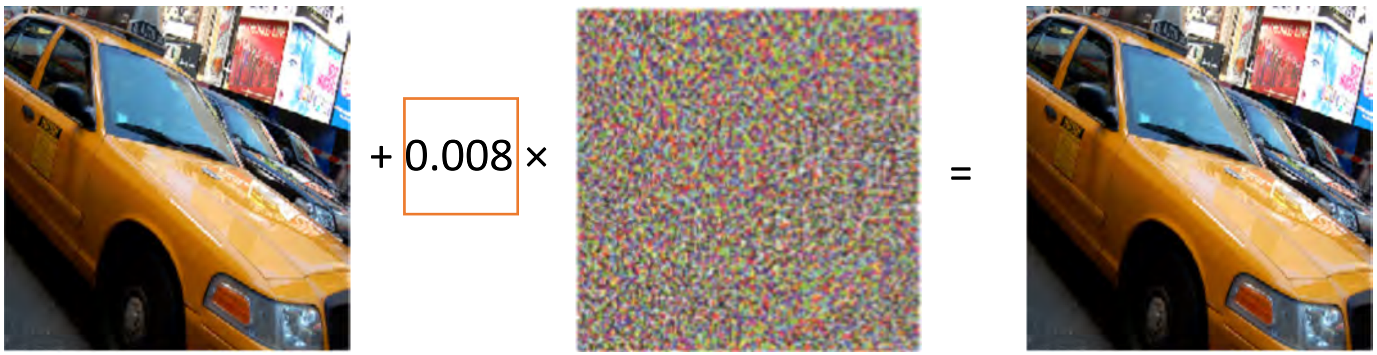
Adversarial Example

Adversarial examples are inputs to machine learning models that an attacker has intentionally designed to cause the model to make a mistake

Prediction: cab

Perturbation noise

Prediction: microwave



confidence 0.895

confidence 0.978

- Input artifacts created from natural data by adding distortions
  - Misclassification
  - Imperceptibility
  - High Confidence

Type of Attack Targets

Adversarial examples are solutions to an optimization problem

- Non-linear and non-convex for many ML models
- No good theoretical tools for describing the solutions to these problems
- Hard to make theoretical argument that a defense will rule out adversarial example

Targeted attack

- Misclassify the predicted class of an input X to an intended target class in Y by crafting the input X via adversarial perturbation.

$$x^{*}: argmin_{x^{*}} L(x, x^{*}) \text{ s.t. } \mathrm{f}(\mathrm{x}^{*}) = y^{*}$$

Untargeted attack

- Misclassify the benign input X by adding adversarial perturbation so that the prediction is changed to some other class Y.

$$x^{*}: -argmin_{x^{*}} L(x, x^{*}) \text{ s.t. } \mathrm{f}(\mathrm{x}^{*}) \neq y$$

Deception with Adversarial Examples

Georgia Tech

- **Untargeted attacks**: Change the true class prediction to any other class prediction

dog

hummingbird

hummingbird

- **Target attacks**: Change the true class prediction to a targeted wrong class prediction

Attack Measurement

Adversarial attack	benign	FGSM	BIM	CW_∞ most	CW_∞ LL	CW_2 most	CW_2 LL	CW_0 most	CW_0 LL	JSMA most	JSMA LL
Adversarial attack image											
prediction	horse	bird	bird	bird	airplane	bird	airplane	bird	airplane	bird	airplane
confidence	0.983	0.725	0.891	1	1	1	1	1	1	0.506	0.384

CIFAR-10		misclassification		mean	DistPerturb			DistPercept		time (s)
attack	UA/TA	ASR	MR	confidence	L_∞	L_2	L_0	SSIM	PSNR	
FGSM	UA	0.85	0.85	0.8647	0.016	0.865	0.996	0.973	48.77	0.021
BIM		0.92	0.92	0.9645	0.008	0.369	0.924	0.995	52.48	0.154
CW_∞	most	1	1	0.9889	0.009	0.326	0.841	0.995	53.25	235.5
	LL	1	1	0.9779	0.014	0.528	0.908	0.989	51.08	243.2
CW_2	most	1	1	0.9867	0.024	0.207	0.428	0.998	55.31	5.772
	LL	1	1	0.9732	0.041	0.357	0.61	0.995	52.81	7.441
CW_0	most	1	1	0.9904	0.574	1.566	0.011	0.962	46.67	355.4
	LL	1	1	0.9757	0.695	2.518	0.024	0.914	44.27	356.7
JSMA	most	1	1	0.5366	0.845	3.739	0.018	0.855	43.19	4.894
	LL	0.99	1	0.392	0.901	5.468	0.036	0.767	40.98	9.858

ImageNet		misclassification		mean	DistPerturb			DistPercept		time (s)
attack	UA/TA	ASR	MR	confidence	L_∞	L_2	L_0	SSIM	PSNR	
FGSM	UA	0.99	0.99	0.641	0.008	3.009	0.982	0.981	43.35	0.019
BIM		1	1	0.997	0.004	1.406	0.854	0.996	46.65	0.185
CW_∞	most	1	1	0.985	0.004	0.915	0.366	0.998	48.61	74.7
	LL	0.95	0.96	0.816	0.01	1.942	0.722	0.993	45.42	237.81
CW_2	most	1	1	0.907	0.009	0.698	0.238	0.999	50.96	13.15
	LL	0.94	0.94	0.777	0.031	1.043	0.313	0.998	48.2	23.13
CW_0	most	1	1	0.97	0.825	4.79	0.001	0.988	41.52	662.7
	LL	1	1	0.806	0.92	9.04	0.005	0.958	38.64	794.9

Two Intriguing Properties of Adversarial Attacks

- Transferability of Adversarial Examples
  - Adversarial examples generated by attacking one ML model often can be effective in attacking other models
    - Ability of an attack against a machine-learning model to be effective against a different, potentially unknown, model
- Divergence of Adversarial Examples
  - Behavior 1: inconsistency in different instances
    - Robustness against adversarial perturbation is different across input instances
  - Behavior 2: inconsistency in different models
    - More than one way to generate successful adversarial example using the same attack method

Characterization-Transferability

Georgia Tech

attack\model		TM	DM 1	DM 2	DM 3	DM 4	DM 5	DM 6	DM 7	Bestκ
FGSM	UA	1	0.224	0.235	0.376	0.436	0.459	0.447	0.422	0.12
BIM	UA	1	0.196	0.207	0.228	0.228	0.261	0.293	0.272	0.1
TFGSM	most	1	0.098	0.104	0.165	0.152	0.154	0.194	0.202	0.072
	LL	1	0	0	0	0.057	0.038	0.075	0.057	0
TBIM	most	1	0.05	0.083	0.141	0.185	0.159	0.176	0.153	0.068
	LL	1	0	0	0	0.003	0.006	0.006	0.003	0
DF	UA	1	0.163	0.224	0.214	0.245	0.276	0.245	0.184	0
CW_∞	most	1	0.06	0.06	0.1	0.13	0.15	0.11	0.11	0
	LL	1	0	0.01	0.02	0.01	0.02	0.03	0.01	0
CW_2	most	1	0.05	0.06	0.1	0.1	0.11	0.1	0.1	0
	LL	1	0	0.01	0.01	0.01	0.02	0.03	0	0
CW_0	most	1	0.11	0.09	0.3	0.23	0.23	0.24	0.25	0.09
	LL	1	0	0	0.08	0.12	0.06	0.1	0.09	0
JSMA	most	1	0.09	0.06	0.13	0.01	0.09	0.1	0.1	0
	LL	1	0	0	0.03	0.05	0.05	0.03	0.02	0
model average		1	0.069	0.076	0.126	0.131	0.139	0.145	0.132	0.03

Transferability in CIFAR-10 models

- Attack Inconsistency: transferability on untargeted attacks stronger than targeted attacks
- Model inconsistency: transferability not equally strong on all models

Characterization-Divergence (untargeted attacks)

Georgia Tech

Different input images of the same class (e.g., digit 1) may have different attack effectiveness even with the same level of noise under the same attack method

For untargeted attacks, consider digit 0 in MNIST

- Instance-level inconsistency

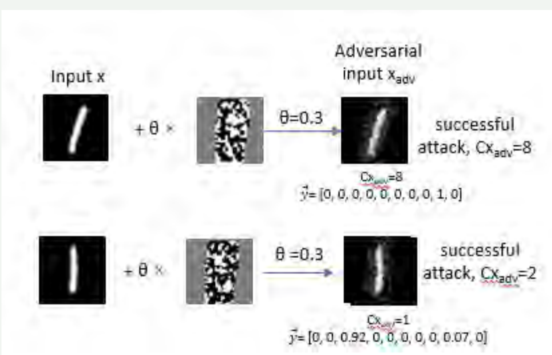


Fig. two inputs draw from the same source class under the FGSM attack.

- Model-level inconsistency

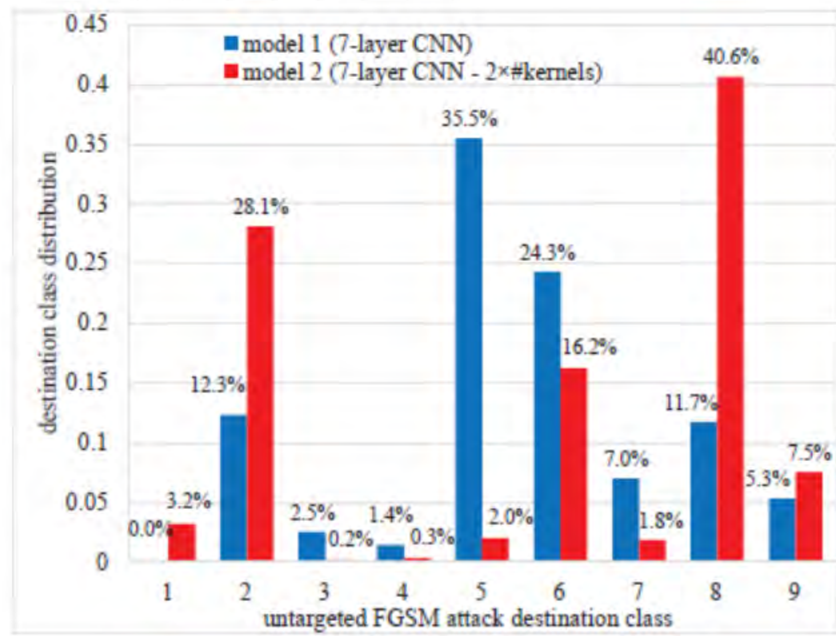


Fig. Destination distributions of 980 images of attacked digit 0. (under FGSM attack)

Characterization-Divergence (targeted attacks)

Georgia Tech

For any benign input, there is more than one way to generate successful adversarial examples (e.g., different θ values) using the same attack method

For targeted attacks, consider digit 4 in MNIST, for two different models under the same attack algorithm

- Instance-level inconsistency
- Model-level inconsistency

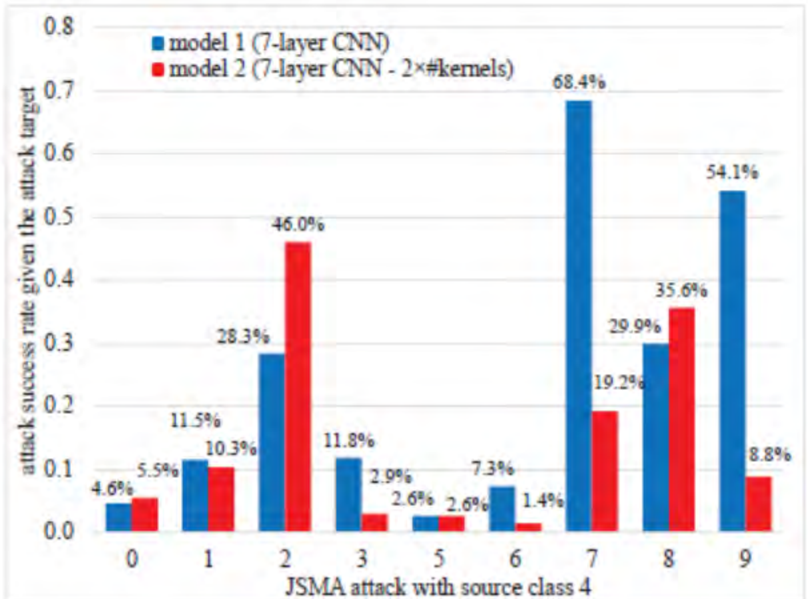
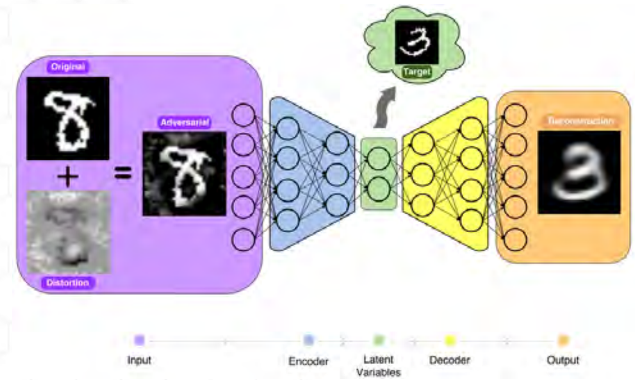


Fig. Attack success rate of 984 images of source digit 4. (under JSMA attack)

Cross-Layer Strategic Ensemble Defense

Cross-layer Ensemble

- Denoising Ensemble: Guarding the input
- Verification Ensemble: Guarding the output



Three Steps

Step 1: Create a pool of candidate ensemble base models

Type 1 structural diversity

Step 2: Create a pool of candidate ensemble teams

Type 2 disagreement diversity

Step 3: Combine, rank and integrate predictions from members of an ensemble committee

Design Principle: Structural Diversity

The structural diversity of DNN models can be achieved by

- Data: varying training dataset
- Structure: different network structure, feature vector size, optimization algorithms, loss function
- Hyperparameters: different mini-batch size, #epochs, #iterations, learning rate functions

model	MNIST	acc	CIFAR-10	acc	ImageNet	acc
TM	CNN1	0.994	DenseNet	0.945	MobileNet	0.695
DM 1	CNN1-$\frac{1}{2}$k	0.986	CNN1	0.78	VGG-16	0.67
DM 2	CNN1-2k	0.995	CNN2	0.746	VGG-19	0.68
DM 3	CNN1-30e	0.988	ResNet-20	0.918	ResNet-50	0.67
DM 4	CNN1-40e	0.988	ResNet-32	0.923	Inception V3	0.735
DM 5	CNN2	0.992	ResNet-44	0.924		
DM 6	CNN2-$\frac{1}{2}$k	0.984	ResNet-56	0.928		
DM 7	CNN2-2k	0.982	ResNet-110	0.926		
DM 8	CNN2-30e	0.984				
DM 9	CNN2-40e	0.986				

Ling Liu, Wenqi Wei, Ka-Ho Chow, Margaret Loper, Mehmet Emre Gursoy, Stacey Truex, and Yanzhao Wu, “Deep Neural Network Ensembles against Deception: Ensemble Diversity, Accuracy and Robustness", The 16th IEEE International Conference on Mobile Ad-Hoc and Smart Systems (IEEE MASS 2019), California, Monterey, Nov 2019.

Design Principle: Disagreement Diversity

The disagreement diversity is measured by the prediction discrepancy of DNN models:

- promote failure independence of ensemble member classifiers
- increase the overall predictive performance (accuracy)
- kappa statistic, Q-statistic, ρ-statistic, and so forth

Pairwise Kappa score:

$$\kappa = \frac{2(N^{11}N^{00} - N^{01}N^{10})}{(N^{11}+N^{10})(N^{01}+N^{00})+(N^{11}+N^{01})(N^{10}+N^{00})}$$

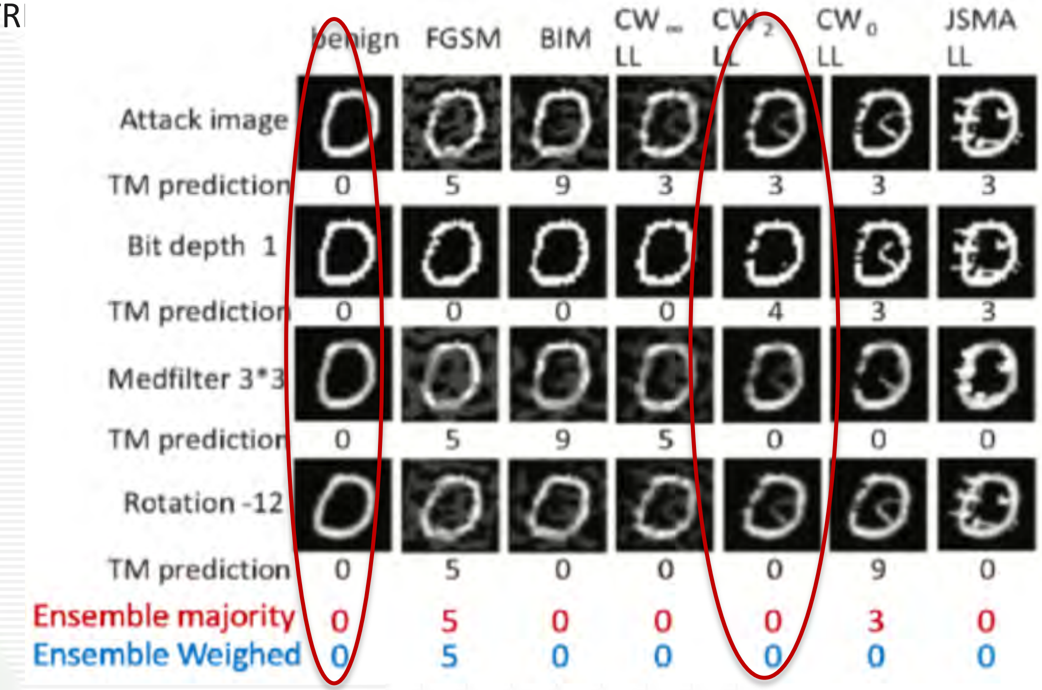
# models	MNIST (TM+)	CIFAR-10 (TM+)	ImageNet (TM+)
3	DM 5,7	DM 2,4	DM 1,3
4	DM 1,5,9	DM 2,4,6	DM 1,3,4
5	DM 1,5,8,9	DM 1,2,4,5	DM 1,2,3,4
6	DM 1,5,6,8,9	DM 1,2,4,5,6	
7	DM 1,4,5,6,8,9	DM 1,2,3,4,5,6	
8	DM 1,4,5,6,7,8,9	DM 1,2,3,4,5,6,7	
9	DM 1,3,4,5,6,7,8,9		
10	DM 1,2,3,4,5,6,7,8,9		

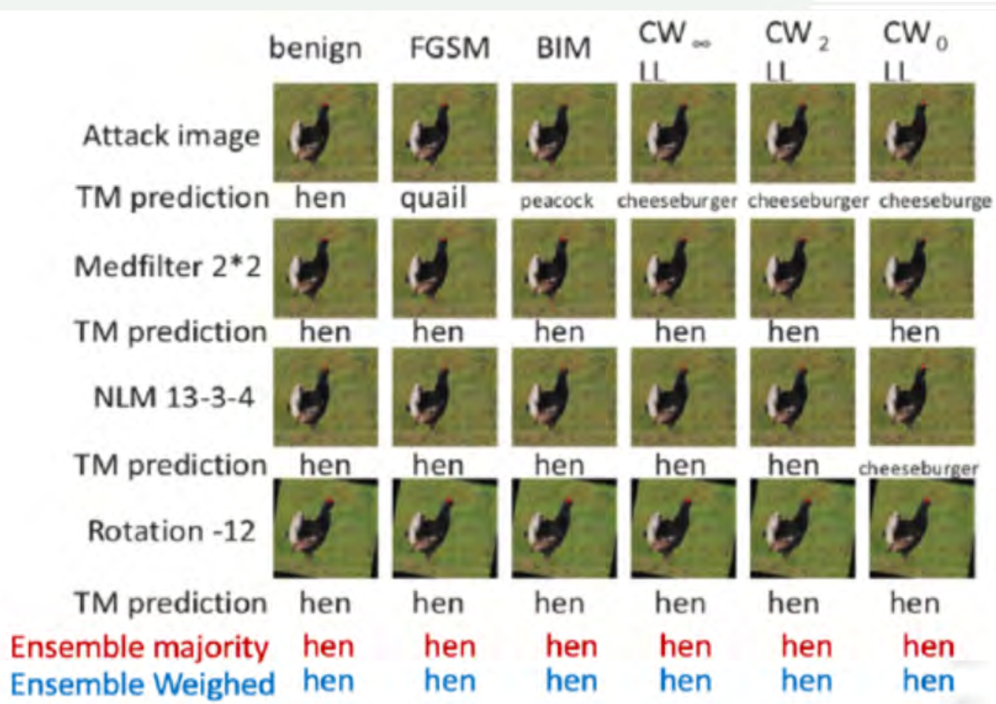
Pairwise Kappa score

2	3	4	5	6	7	8
0.148	0.17	0.204	0.182	0.214	0.217	0.252
1	0.677	0.562	0.507	0.507	0.54	0.551
	1	0.564	0.475	0.496	0.485	0.54
		1	0.617	0.57	0.594	0.672
			1	0.64	0.63	0.72
				1	0.641	0.661
					1	0.72
						1

GT/GTRI

Input Diversity Ensemble





Cross-Layer Strategic Ensemble Defense

We compare three different output-level ensembles:

- Randbase: random models from the baseline model pool,
- Rand-kappa: ensemble randomly selected in the kappa ensemble model list,
- Best-kappa: ensemble in the kappa ensemble model list with the best auto-repairing/auto-flagging performance.

Measured in defense success rate = repair success rate + flag success rate

	model	benign acc	FGSM	BIM	TFGSM		TBIM		CW_∞		CW_2		CW_0		average
			UA		most	LL	most	LL	most	LL	most	LL	most	LL	
ImageNet	TM	0.695	0.01	0	0	0.09	0	0.21	0	0.04	0	0.06	0	0	0.034
	DM 1	0.67	0.73	0.77	0.73	0.77	0.82	0.82	0.81	0.81	0.81	0.82	0.8	0.79	0.79
	DM 2	0.68	0.7	0.78	0.72	0.76	0.81	0.85	0.83	0.83	0.84	0.84	0.81	0.76	0.794
	DM 3	0.67	0.78	0.84	0.8	0.81	0.84	0.84	0.85	0.83	0.83	0.84	0.83	0.8	0.824
	DM 4	0.735	0.86	0.85	0.87	0.88	0.91	0.93	0.92	0.91	0.92	0.9	0.91	0.84	0.892
	RandBase: DM 1,2,3	0.770	0.92	0.92	0.92	**0.97**	0.91	0.93	0.92	0.94	0.91	0.93	0.92	0.95	0.928
	Randκ: DM 1,2,3,4	0.755	0.83	0.9	0.85	0.87	0.92	0.92	0.92	0.91	0.92	0.9	0.89	0.89	0.893
	Bestκ: DM 1,3,4	0.805	**0.94**	0.93	0.95	**0.97**	0.97	0.95	0.96	0.96	**0.95**	0.95	0.91	0.97	0.951
	rot_6 → RandBase	0.785	0.93	0.9	0.87	0.94	0.92	0.95	0.93	0.94	0.91	0.95	0.89	0.94	0.923
	rot_6 → Randκ	0.745	0.85	0.87	0.83	0.85	0.88	0.87	0.87	0.85	0.86	0.88	0.89	0.88	0.865
	rot_6 → Bestκ	0.825	0.89	0.94	0.87	0.95	0.96	0.96	0.96	0.93	0.93	**0.96**	0.96	**0.98**	0.941
	rot_6 + Bestκ	0.89	0.89	**0.96**	**0.99**	0.92	**1**	**1**	**1**	**1**	0.93	**0.96**	**0.99**	0.97	**0.97**

Output Diversity Ensemble

benign | FGSM | BIM | CW $_\infty$ LL | CW $_2$ LL | CW $_0$ LL | JSMA LL

dog | ship | ship | deer | deer | deer | deer

TM	dog (0.995)	ship (0.998)	ship (1)	deer (0.97)	deer (0.905)	deer (0.973)	deer (0.3)
DM 1	dog (0.999)	dog (0.987)	dog (0.999)	dog (0.999)	dog (0.999)	dog (0.901)	dog (0.834)
DM 2	dog (1)	dog (1)	dog (1)	dog (1)	dog (1)	dog (1)	dog (0.815)
DM 3	dog (0.999)	ship (0.782)	dog (0.582)	dog (0.992)	dog (0.997)	dog (0.84)	frog (0.687)
DM 4	dog (1)	dog (0.998)	dog (1)	dog (1)	dog (1)	dog (0.999)	frog (0.592)
DM 5	dog (1)	dog (0.953)	dog (1)	dog (1)	dog (1)	deer (0.571)	deer (0.827)
DM 6	dog (1)	dog (0.679)	dog (0.995)	dog (1)	dog (1)	dog (0.999)	frog (0.592)
DM 7	dog (0.999)	ship (0.897)	dog (0.488)	dog (0.985)	dog (0.981)	dog (0.936)	ship (0.442)
RandBase3	dog (1)	dog (1)	dog (1)	dog (1)	dog (1)	dog (1)	dog (0.667)
RandBase5	dog (1)	dog (0.6)	dog (1)	dog (1)	dog (1)	dog (1)	frog (0.6)

Ongoing Research on Adversarial Learning for Robust AI / ML

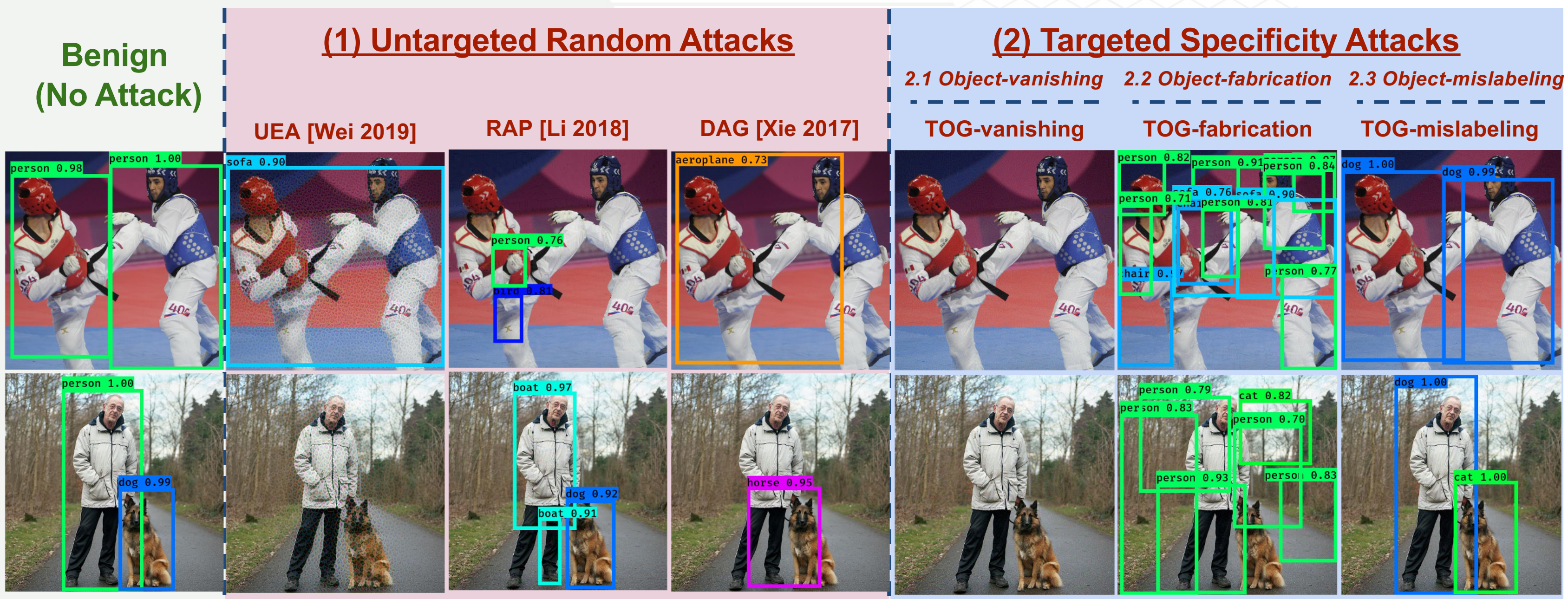
In addition to Adversarial input attacks to single task learners and our cross-layer ensemble mitigation, we are:

- Working on risk assessment and risk mitigation for real time object detection systems
- Developing risk mitigation algorithms using high diversity ensemble training

What Malicious Effects can be Caused by the Attack?

Each of the three tasks (existence, bounding box, and class label) can be considered as an attack surface

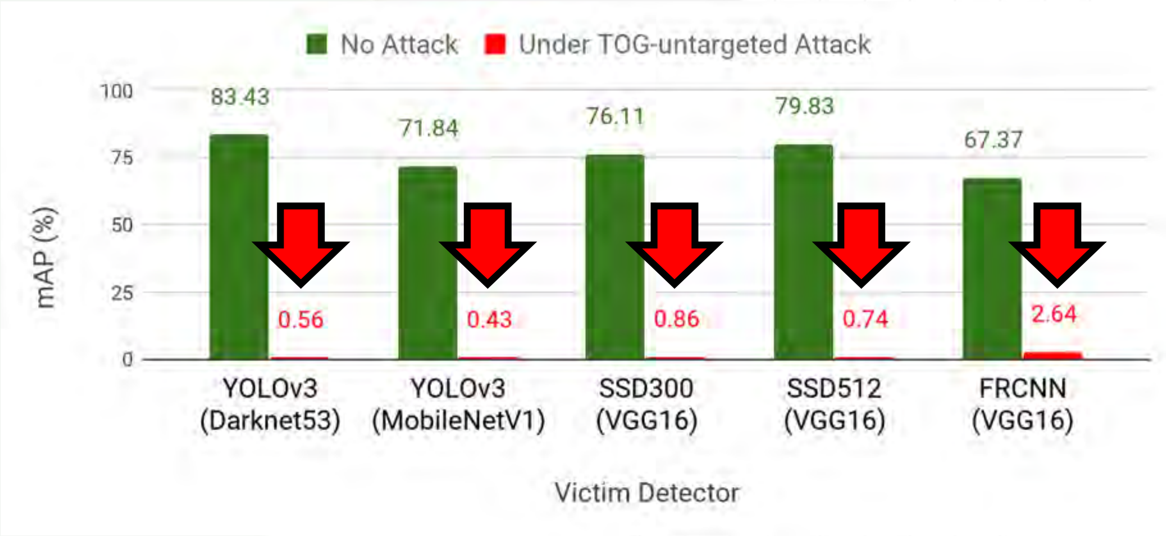
⇒ Attack effects can be more **flexible** than simply misclassification!



Vulnerabilities of Popular Deep Object Detectors

The TOG attack can severely damage any state-of-the-art deep object detectors

- 67.37~83.43% ⇒ 0.56~2.64% = NO detection capability



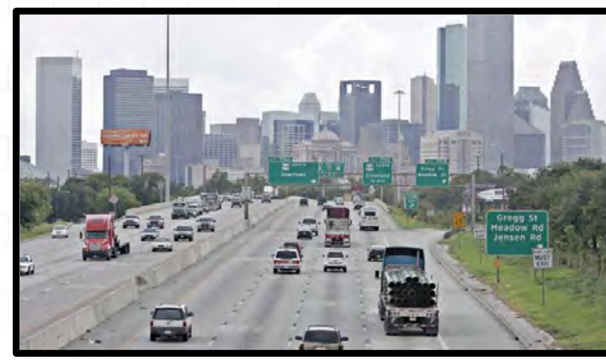
TOG <u>targeted</u> attacks (object-vanishing, object-fabrication, object-mislabeling) are equally detrimental!!!

Future Applications: Edge Intelligence in Cities

- Edge Intelligence: data is collected, analyzed, and insights produced near the end user
- Goal: provide real-time information to the user via computing, network optimization, and multi-tier AI
- Many applications use video cameras as edge device

Self-Driving Vehicles

Public Safety



Transportation Systems

More complicated attack - more complicated defense!

Concluding Remark

- Adversarial Attacks are serious deception threats with two intriguing properties
  - Transferability
  - Divergence
- Cross-Layer Strategic Ensembles can be an effective defense against deception
  - Quantifying ensemble diversity and guaranteeing ensemble robustness → Technical Challenges
- Future research
  - New generations of attack and defense methods for comparison
  - Application of ensembles to video attacks

Wei W, Liu L, Loper M, Chow KH, Gursoy E, Truex S, Wu Y. “A Framework for Evaluating Gradient Leakage Attacks in Federated Learning”, 25th European Symposium on Research in Computer Security (ESORICS 2020), September 2020.

Wei W, Liu L, Loper M, Chow KH, Gursoy E, Truex S, Wu Y. “Cross-Layer Strategic Ensemble Defense Against Adversarial Examples”, 2020 International Conference on Computing, Networking and Communications (ICNC), February 2020.

Thank you!

Q&A