Adversarial Examples

CSIT375/975 AI and Cybersecurity

Dr Wei Zong

SCIT University of Wollongong

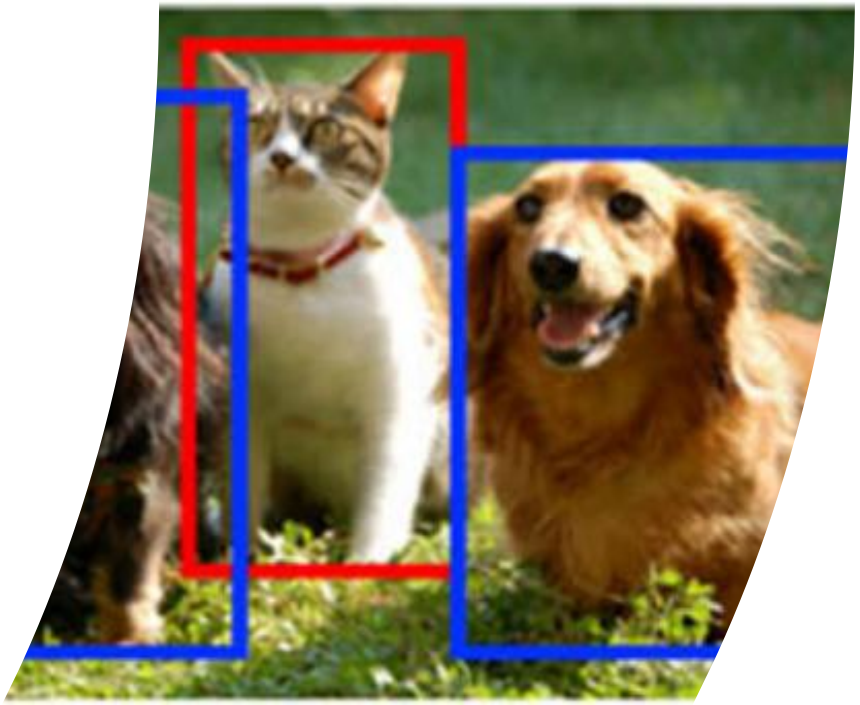
Disclaimer: The presentation materials come from various sources. For further information, check the references section

Outline

- Adversarial machine learning
- Adversarial examples
  - White-box attacks
  - Black-box attacks
  - Physical attacks
  - Universal adversarial perturbations
  - Unrestricted attacks

Machine learning models are doing well

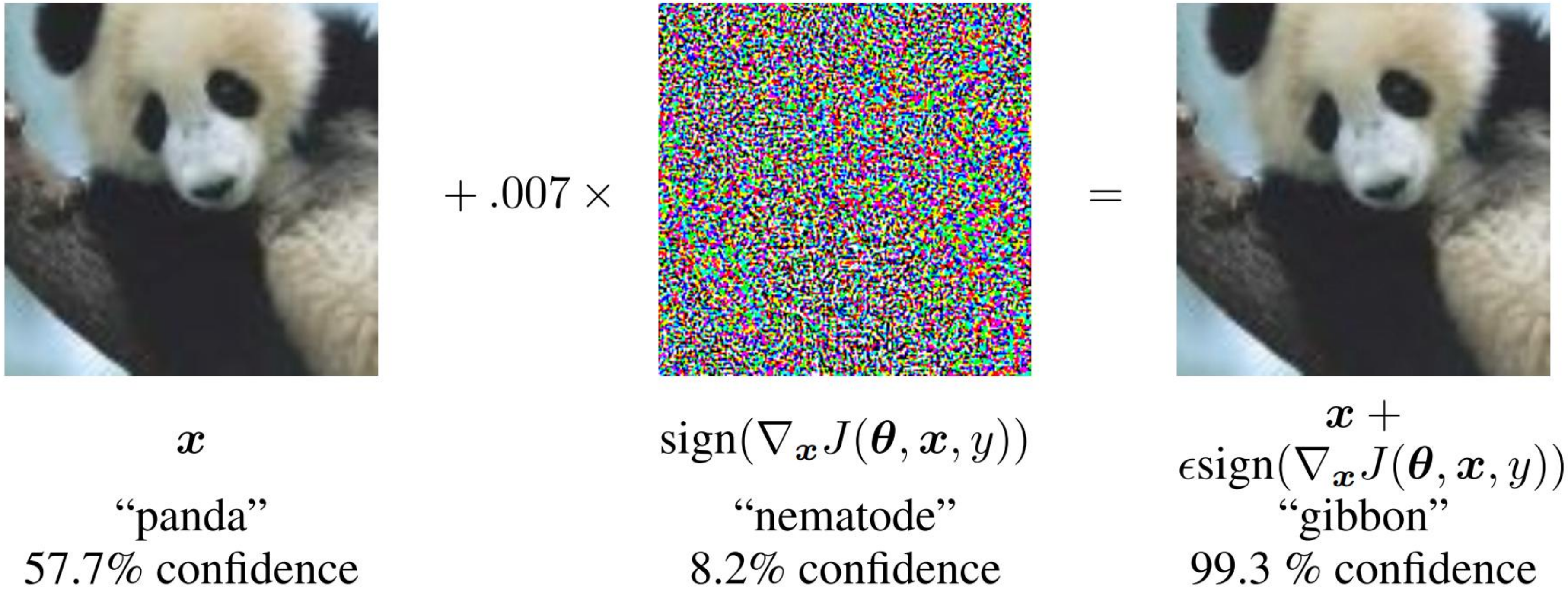
- Image classification
- Object detection
- Self-driving cars
- Speech-to-text
- Etc ...



Are models just Clever Hans?

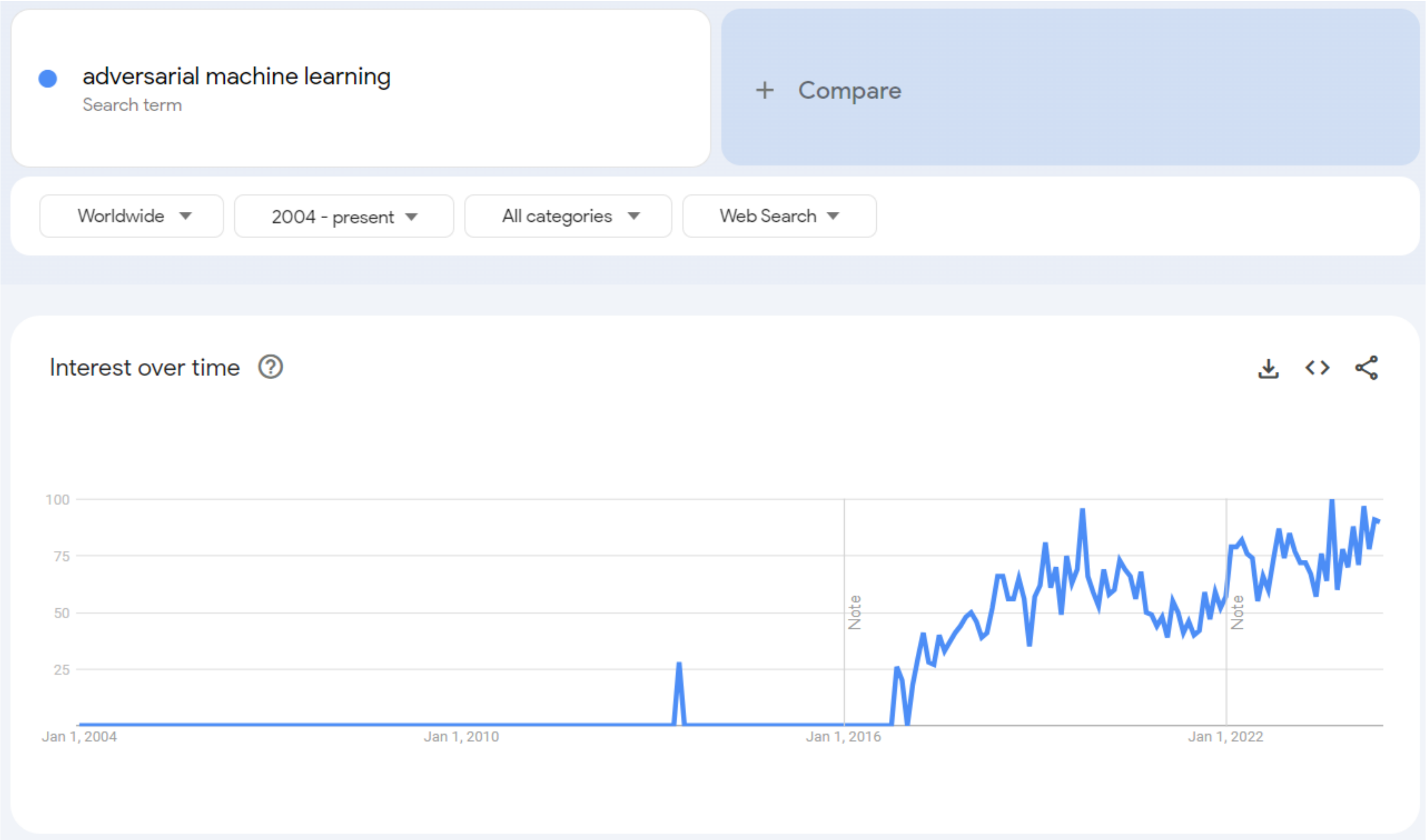
Clever Hans performing in 1904 (https://en.wikipedia.org/wiki/Clever_Hans)

Are models just Clever Hans?


$+ .007 \times$
$=$
$\boldsymbol{x}$
"panda"
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$\boldsymbol{x} +$
$\epsilon \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

Explaining and Harnessing Adversarial Examples. ICLR (Poster) 2015

Adversarial Machine Learning



● adversarial machine learning

Search term

Worldwide

2004 - present

All categories

Web Search

Interest over time

https://trends.google.com/

What is Adversarial Machine Learning?

Attackers aware of the ML techniques being deployed against them can, for example, contaminate the training data to manipulate a learned ML classifier in order to evade subsequent classification, or can manipulate the specific metadata upon which the ML algorithms make their decisions and exploit identified weaknesses in these algorithms so called **Adversarial Machine Learning**.

https://www.dst.defence.gov.au/publication/adversarial-machine-learning-cyber-security-ngtf-project-scoping-study

Outline

- Adversarial machine learning

- Adversarial examples
  - White-box attacks
  - Black-box attacks
  - Physical attacks
  - Universal adversarial perturbations
  - Unrestricted attacks

Adversarial Examples

- Definition
  - Given original input x, an adversarial example x’ is obtained by slightly modifying x.
    - x’ ← x + δ
    - x’ ≈ x (e.g., $||x' - x||_1 = \sum |(x' - x)_i| < \epsilon$)
    - The vector norm $||v||_p$, also called ***p-norm*** *or* $\boldsymbol{L_p}$ **norm,** for p=1,2,4, …, is defined as:

$$||v||_p = \left(\sum_i |v_i|^p\right)^{\frac{1}{p}}$$

      - $||v||_\infty = \max_i |v_i|$: maximum absolute value.
      - p=1: sum of absolute values.
      - p=2: the Euclidean distance.

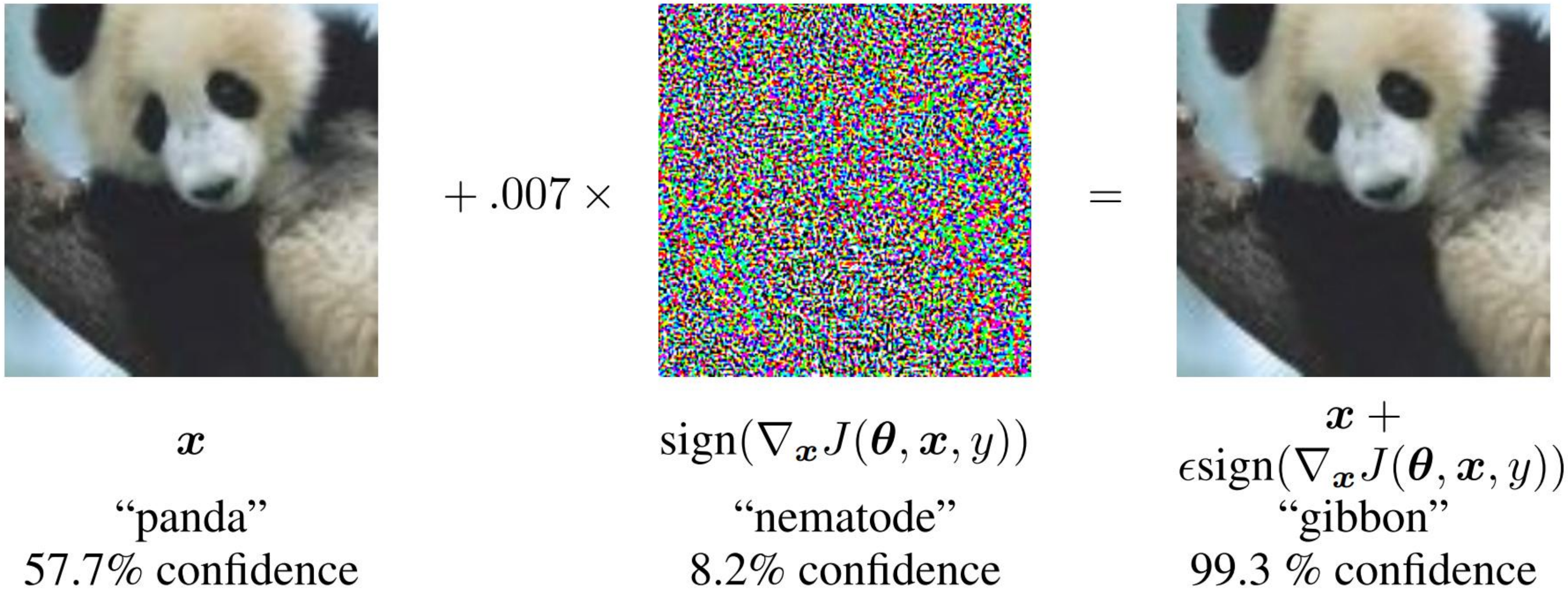
- Attacks
  - Targeted attack vs. untargeted attack
    - Targeted: fool a model to output predefined labels.
    - Untargeted: fool a model to output incorrect labels.

Fast Gradient Signed Method (FGSM)

- White-box attack
  - An adversary (attacker) knows **everything** about the target model.
    - Model architecture
    - Model weights
    - Training data
    - Etc.
  - An adversary can calculate gradients with respect to input data.
    - For constructing adversarial examples.
- FGSM: fool a model by calculating gradients only once.
  - Very **fast**.
  - Simple but effective.
- Let J(θ, x, y) be the cost function used to train a model (a deep neural network):
  - $\delta \leftarrow \epsilon \cdot sign\left(\nabla_x J(\theta, x, y)\right)$ // Sign of gradients; ϵ is a small value for imperceptibility.
  - $x' \leftarrow x + \delta$ // "**+**" for **untargeted** attack, if **y is the ground truth label**.
    // "**-**" for **targeted** attack, if **y is NOT the ground truth label**.
  - $x' \leftarrow clip(x', 0, 1)$ // (pixels range from [0, 1]).

Fast Gradient Signed Method (FGSM)

- Let J(θ, x, y) be the cost function used to train a model (y is the ground truth label):
  - $\delta \leftarrow \epsilon \cdot sign\left(\nabla_x J(\theta, x, y)\right)$
  - $x' \leftarrow x + \delta$
  - $x' \leftarrow clip(x', 0, 1)$

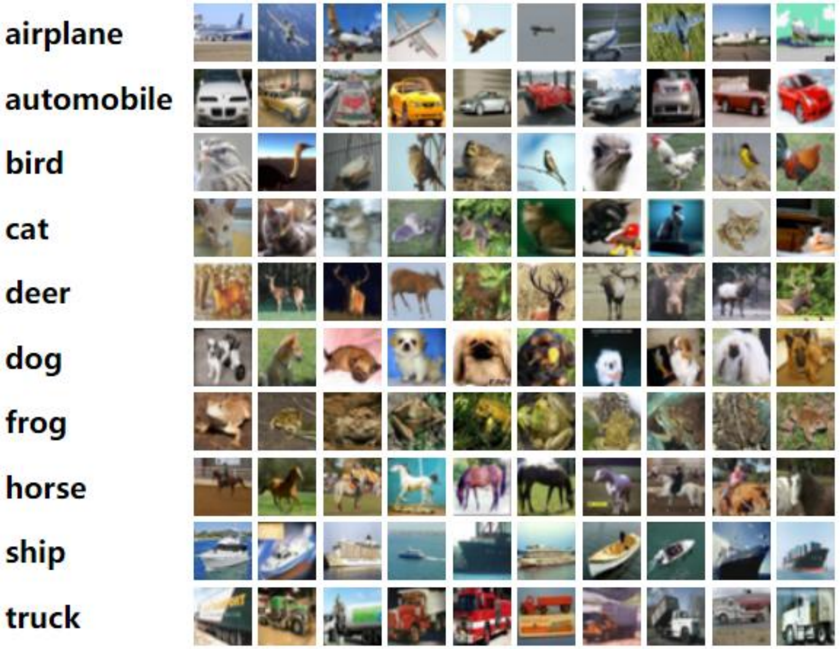

$+ .007 \times$
$=$
$\boldsymbol{x}$
"panda"
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$\boldsymbol{x} +$
$\epsilon \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

Fast Gradient Signed Method (FGSM)

MNIST

- $\epsilon = 0.25$
- an error rate of 99.9% with an average confidence of 79.3%

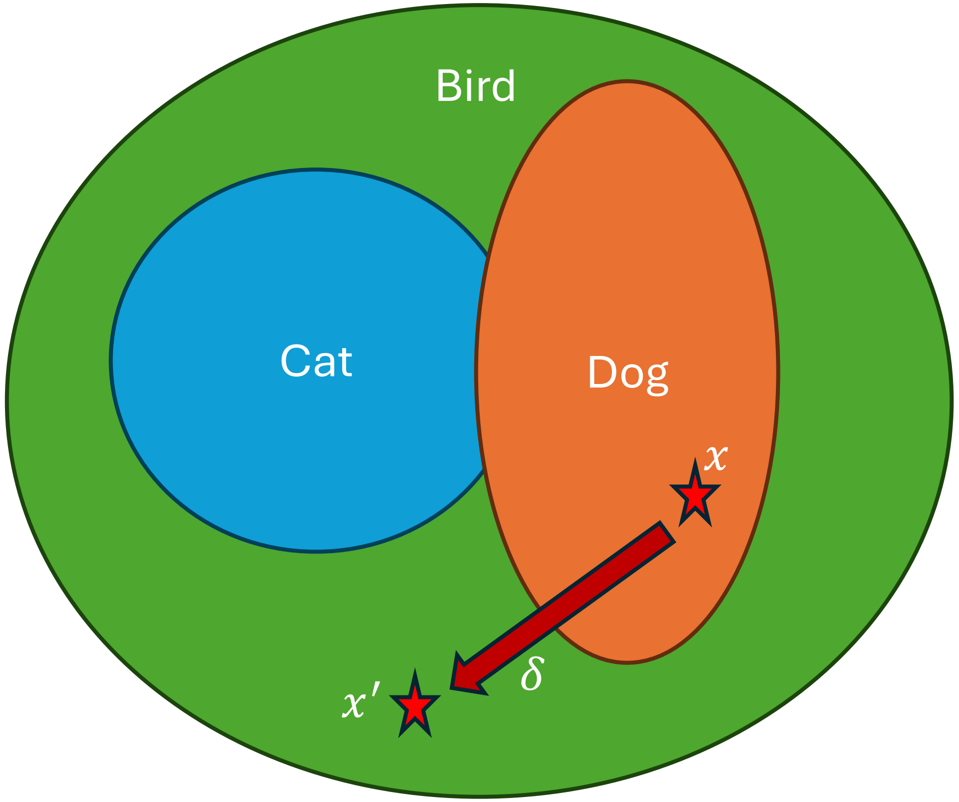
CIFAR10



- $\epsilon = 0.1$
- an error rate of 87.15% with an average confidence of 96.6%

Fast Gradient Signed Method (FGSM)

- **May** fail to produce **targeted** attacks
  - Fool a model to output predefined labels.
  - May not fool a model to classify a dog as a cat.



Projected Gradient Descent (PGD)

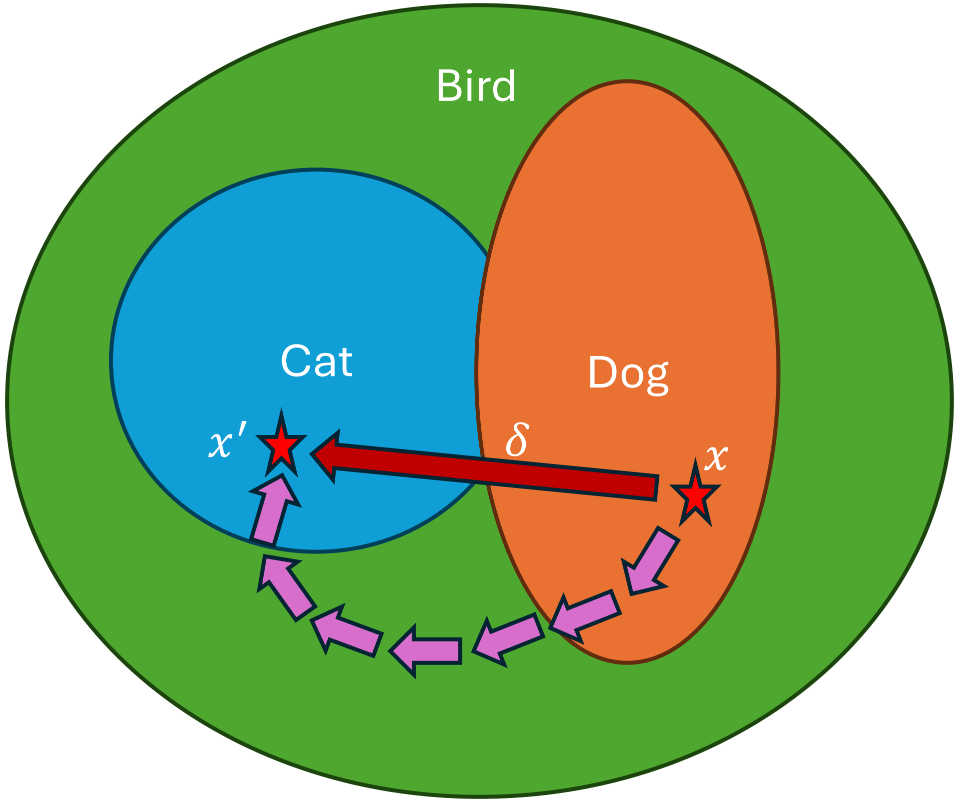
- Much stronger white-box attack than FGSM
- High performance for **targeted** attacks
  - Fool a model to predict predefined labels.

- Key idea: a single step may fail, so we do **multiple** steps!

1. $x' \leftarrow x$
2. $x' \leftarrow x' - \epsilon \cdot sign\left(\nabla_{x'} J(\theta, x', y)\right)$

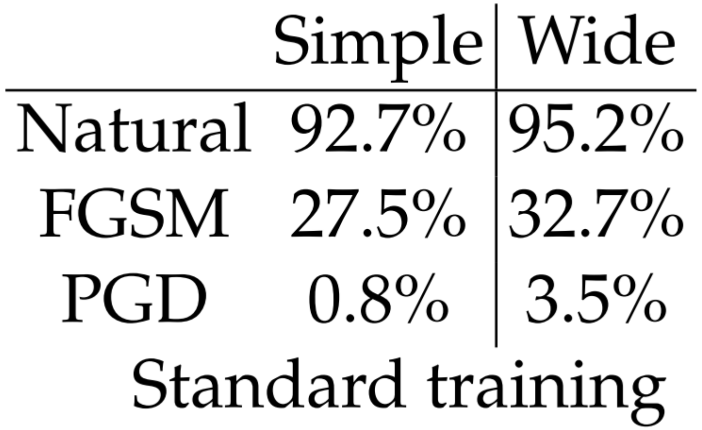
   (**minus for targeted attack; y is the target label**)
3. $\delta \leftarrow clip\left(||x' - x||_p, -\epsilon, \epsilon\right)$

   (**require** $||\delta||_p \leq \epsilon$)
4. $\mathrm{x}' \leftarrow clip(x + \delta, 0, 1)$
5. go back to Step 2 if x' is not predicted as target.



Projected Gradient Descent (PGD)

- Much stronger white-box attack than FGSM
  - Can also generate untargeted attacks
    - CIFAR10, $\epsilon = 0.03$



	Simple	Wide
Natural	92.7%	95.2%
FGSM	27.5%	32.7%
PGD	0.8%	3.5%

Standard training

Carlini-Wagner (CW) Attack

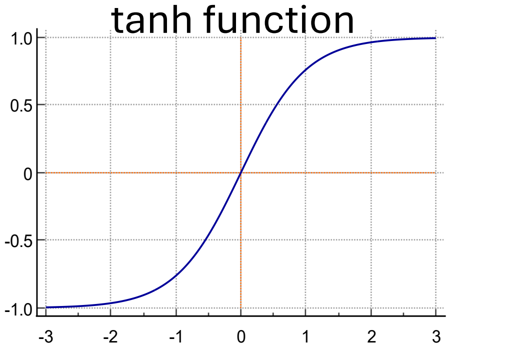
- Another strong and widely used white-box attack
  - Focus on L_2 version
    - Finding adversarial examples that will have low distortion in the L_2 metric.
    - CW attack has L_0 and L_∞ versions.

- Key idea for **targeted** attack: let the target logit value be larger than the second largest one by a specific gap.
  - Do not need to call loss function for training, i.e., cross-entropy loss.
    - Different from FGSM and PGD
  - The gap is used to control for the prediction confidence.
    - Prediction is more confident as the gap becomes larger.
    - Also introduce more distortions.
  - Straightforward to adjust for **untargeted** attack.
    - Let the correct logit value be smaller than the largest value by a gap.

Carlini-Wagner (CW) Attack

- L_2 targeted attack optimization goal:

$$\underset{w}{\text{minimize}} \quad \|\frac{1}{2}(\tanh(w)+1)-x\|_2^2 + c\cdot f(\frac{1}{2}(\tanh(w)+1)$$



https://www.medcalc.org/manual/tanh-function.php

  - tanh: hyperbolic tangent function:
    - Convert a real number to (-1, 1)
  - $\frac{1}{2}(\tanh(w)+1)$ represents the adversarial example.
    - In the range (0, 1)
  - f calculates the gap between target logit value and the largest one:

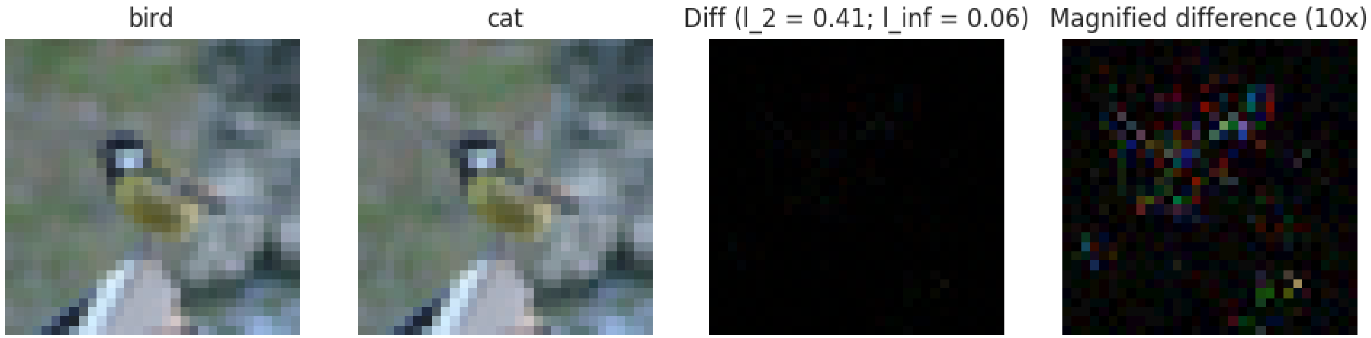
$$f(x') = \max(\max\{Z(x')_i : i \neq t\} - Z(x')_t, -\kappa)$$

    - $Z(\cdot)_i$ denotes the i^{th} logit value.
    - t denotes the target label.
    - k specifies the gap.
      - Larger k indicates larger prediction confidence.
  - c balances the imperceptibility and fooling power.

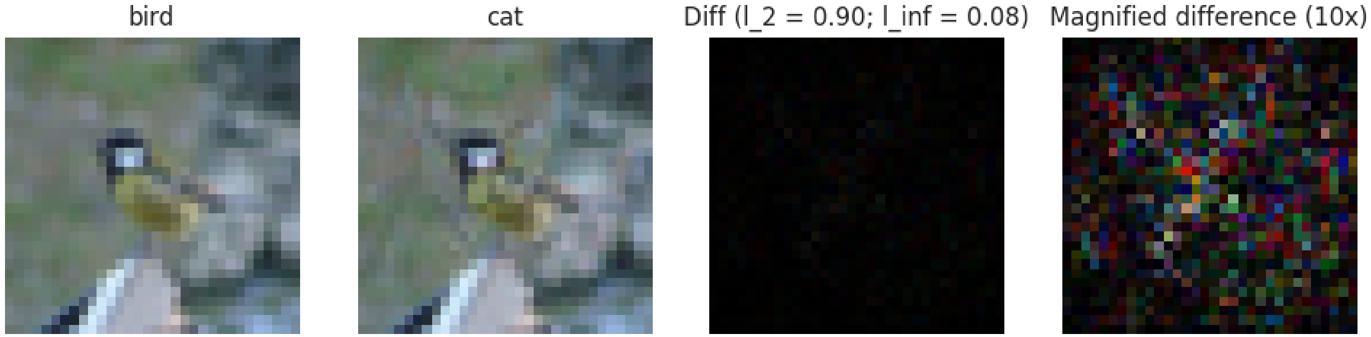
Carlini-Wagner (CW) Attack

- Results (lab2)

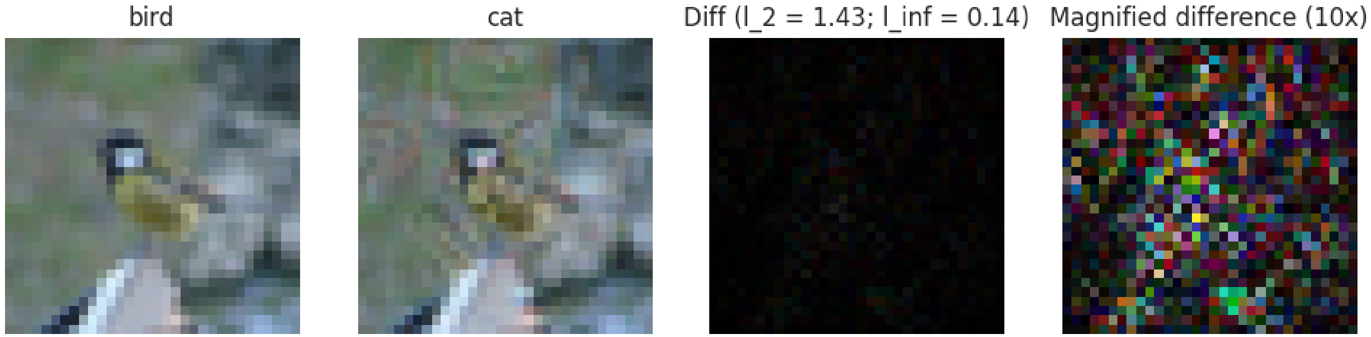
$k = 1$



$k = 40$



$k = 80$



Grey-box & Black-box Adversarial Examples

- Grey-box attacks.
  - Stricter assumptions than white-box attacks.
  - Partial knowledge of the target model
    - May know the architecture of the target model
    - May know the logits output from the target model (in addition to predicted labels).
    - Etc.
  - Cannot compute gradients w.r.t input data.

- Black-box attacks
  - The most practical assumptions.
  - An adversary only observes input and output pairs.
    - No knowledge about the internal workings of the target model.
  - Cannot compute gradients w.r.t input data.

Black-box Adversarial Examples

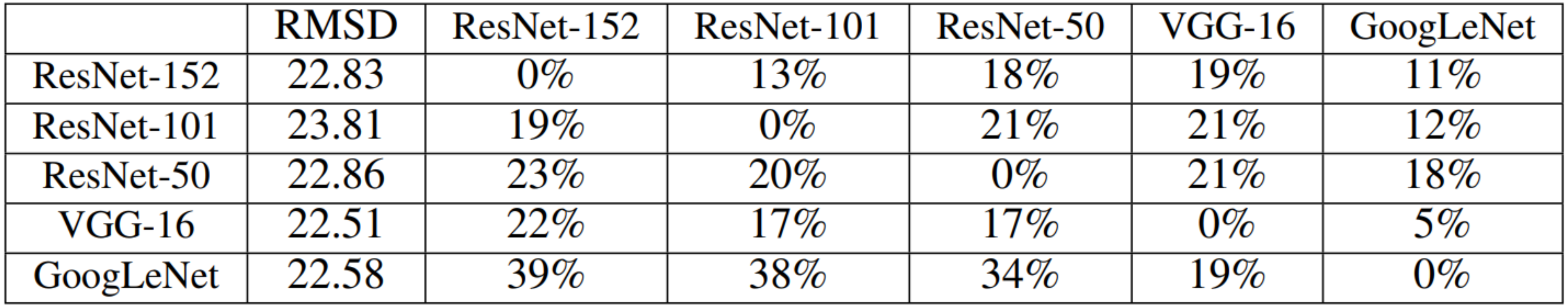
- Attacking strategy
  - Train a **similar** model on **similar** data compared to the target model.
    - It would be ideal if training the **same** model on the **same** set of data.
  - Generate white-box attacks on this model.
  - Evaluate these attacks on the target model.
    - It’s not guaranteed that the attacks will succeed.

- Intuition: **similar** models trained on **similar** data would be **similar**.
  - Generate attacks using a model with **white-box** access.
  - Evaluate these attacks on the target **black-box** model.

Black-box Adversarial Examples

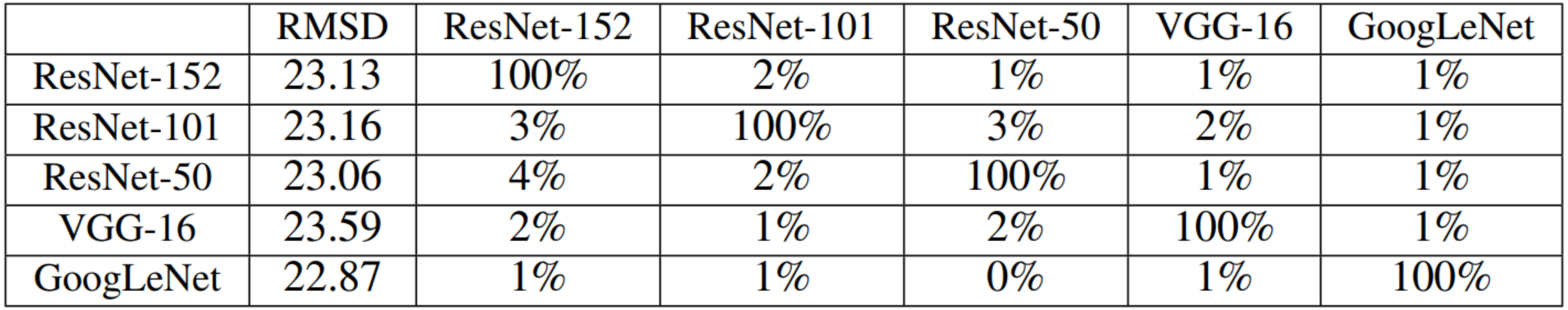
- **Single** model approach
  - Generate attacks using a **single** model.
  - Evaluate these attacks on the target model.

Accuracy of Untargeted Attacks



	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%

Matching Rate of Targeted Attacks



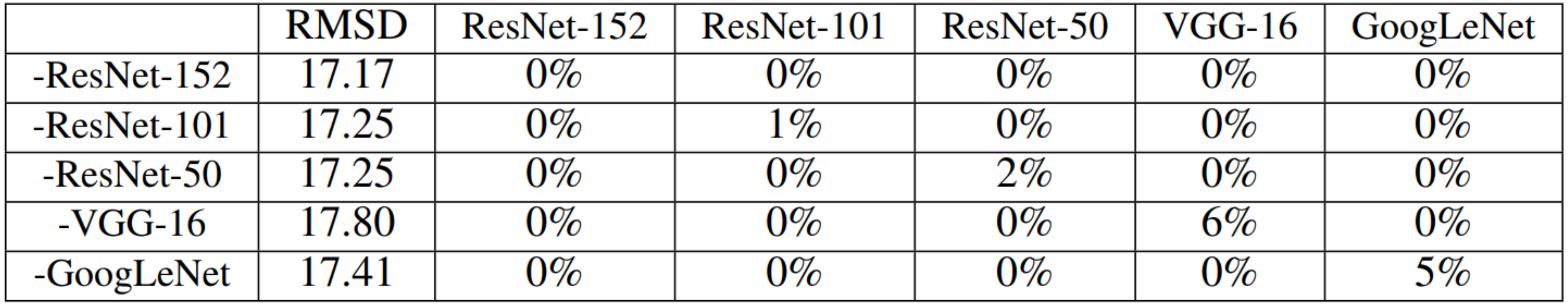
	RMSD	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
ResNet-152	23.13	100%	2%	1%	1%	1%
ResNet-101	23.16	3%	100%	3%	2%	1%
ResNet-50	23.06	4%	2%	100%	1%	1%
VGG-16	23.59	2%	1%	2%	100%	1%
GoogLeNet	22.87	1%	1%	0%	1%	100%

Liu, Y., Chen, X., Liu, C. and Song, D., 2016. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*.

Black-box Adversarial Examples

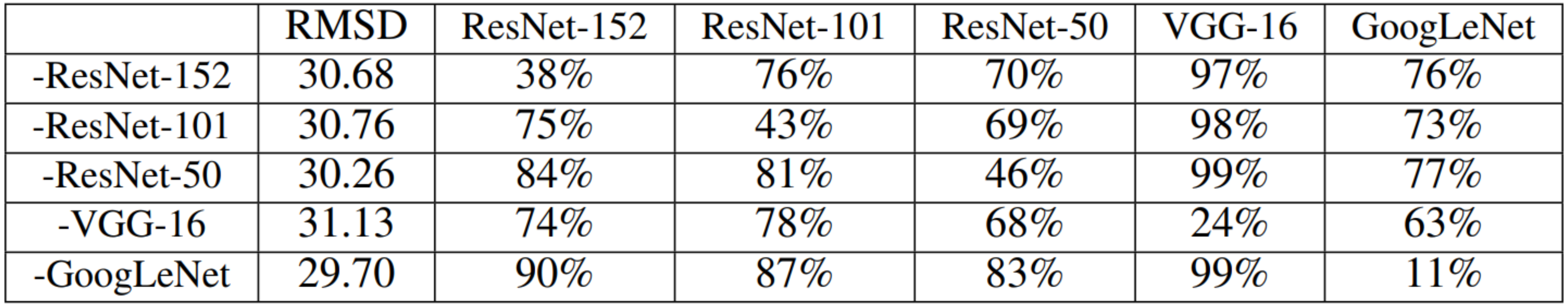
- **Ensemble** approach
  - Generate attacks using an **ensemble** of models.
  - Evaluate these attacks on the target model.

Accuracy of Ensemble Untargeted Attacks



	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%

Matching rate of Ensemble Targeted Attacks

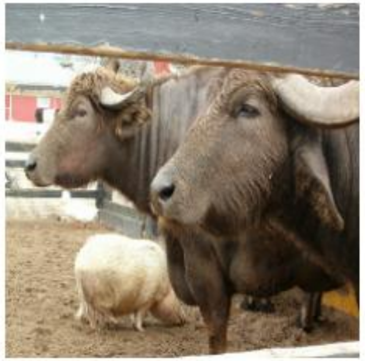
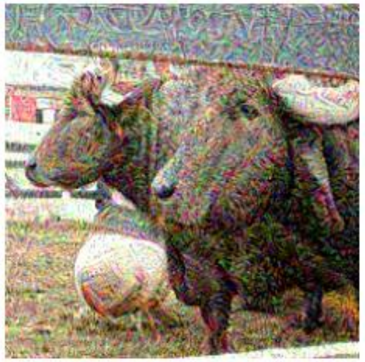
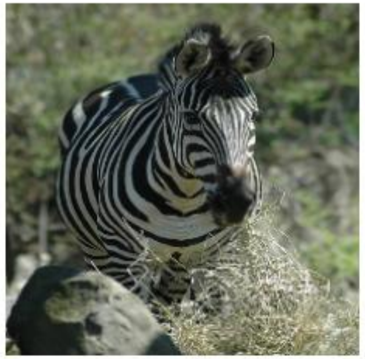
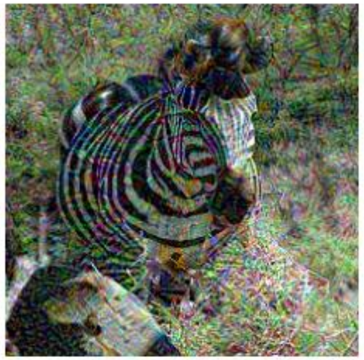


	RMSD	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
-ResNet-152	30.68	38%	76%	70%	97%	76%
-ResNet-101	30.76	75%	43%	69%	98%	73%
-ResNet-50	30.26	84%	81%	46%	99%	77%
-VGG-16	31.13	74%	78%	68%	24%	63%
-GoogLeNet	29.70	90%	87%	83%	99%	11%

Liu, Y., Chen, X., Liu, C. and Song, D., 2016. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*.

Black-box Adversarial Examples

- Examples

	water buffalo, water ox, Asiatic buffalo, Bubalus bubalis	herd, milk, beef cattle, farmland, cow	rugby ball		pastime, print, illustration, art
	zebra	equid, stripe, savanna, zebra, safari	apiary, bee house		wood, people, outdoors, nature

Liu, Y., Chen, X., Liu, C. and Song, D., 2016. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*.

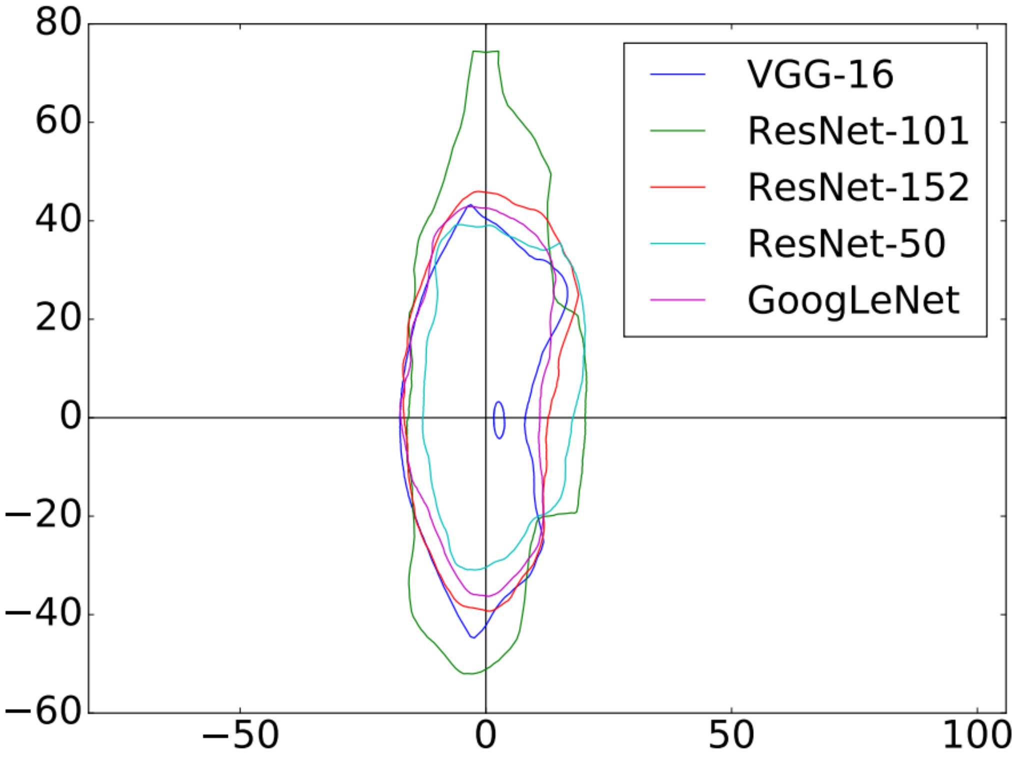
Black-box Audio Adversarial Examples

- Hard to hide noise in images.
- But we can hide noise in music!
  - Using a surrogate model to generate attacks.
  - Transfer to black-box commercial models.
  - https://sites.google.com/view/devil-whisper

Black-box Adversarial Examples

anemone fish

- Visualizing decision boundaries for correct predictions.
  - All points in the decision boundaries are classified as the ground truth label.
    - X-axis: the gradient direction of VGG-16
    - Y-axis: a random orthogonal direction
    - The origin of the coordinate plane corresponds to the original image.
    - The units of both axes are 1-pixel values.
  - Different models share similar decision boundaries.
    - Can explain the high performance of ensemble untargeted attacks.

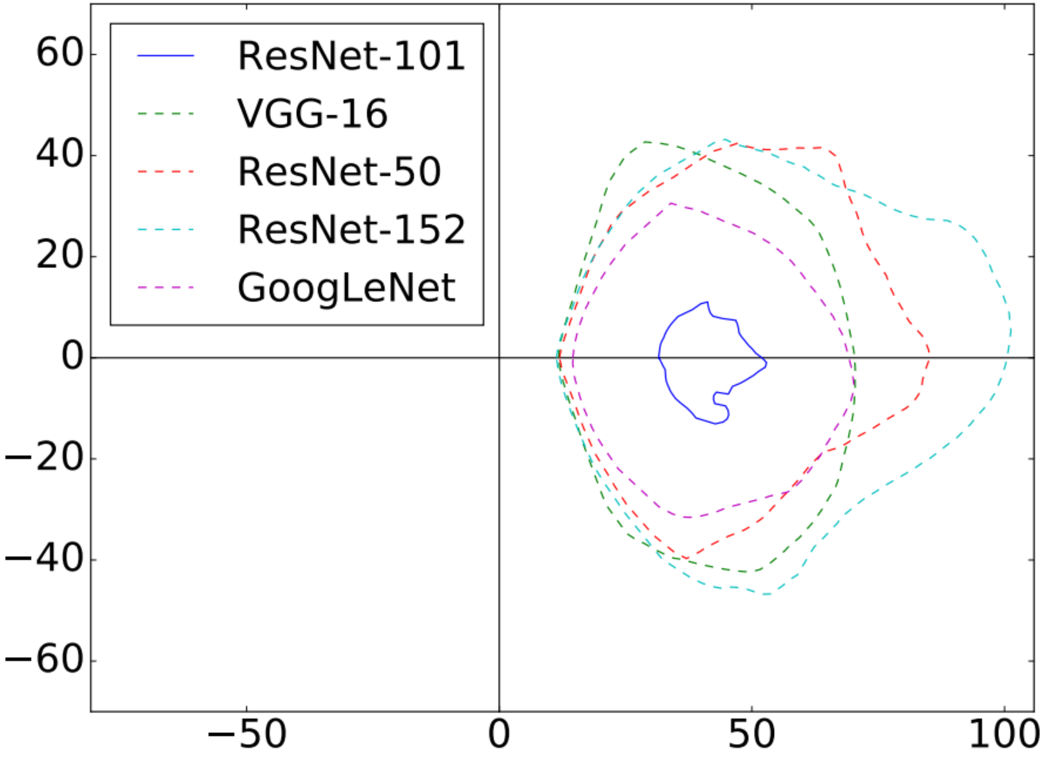


Liu, Y., Chen, X., Liu, C. and Song, D., 2016. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*.

Black-box Adversarial Examples

- Visualizing decision boundaries for targeted attack.
  - All points within the decision boundaries are classified as the target label (i.e., window).
    - X-axis: the targeted adversarial direction
      - The targeted adversarial direction is computed as the difference between the original image and the adversarial image.
    - Y-axis: a random orthogonal direction.
    - The units of both axes are 1-pixel values.
    - The origin of the coordinate plane corresponds to the original image.
  - The ensemble attack contains all models except ResNet101.
    - Different models share similar decision boundaries for the target label.
    - Can explain the success of ensemble targeted attacks.
      - The probability of fooling ResNet101 is not small.

anemone fish



Liu, Y., Chen, X., Liu, C. and Song, D., 2016. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*.

Physical Adversarial Examples

- Machine learning / deep learning models are widely deployed in real-world applications.
  - Self-driving cars.

https://www.wired.com/story/self-driving-cars-are-being-put-on-a-data-diet/

- Adversarial examples in the physical world.
  - Techniques discussed so far focus only on digital attacks.
    - Files are transferred in a digital format.
  - Physical attacks are also possible.
    - Attacks are captured by physical devices, e.g., cameras.
    - Can cause serious harm, e.g., loss of Life.
    - Pose practical threats to the public.

Physical Adversarial Examples

- Physical World Challenges
  - Environmental Conditions.
    - The distance and angle of a camera in an autonomous vehicle with respect to a road sign varies continuously.
    - Other environmental factors include changes in lighting/weather conditions, and etc.
  - Spatial Constraints.
    - a physical road sign, the attacker cannot manipulate background imagery.
  - Physical Limits on Imperceptibility.
    - Transferring minute perturbations to the real world, we must ensure that a camera is able to perceive the perturbations.
  - Fabrication Error.
    - A fabrication device, such as a printer, can produce some color error.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T. and Song, D., 2018. Robust physical-world attacks on deep learning visual classification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1625-1634).

Physical Adversarial Examples

- Robust Physical Perturbation
  - Targeted attack
  - **Key idea**: including all potential transformations into the calculation of perturbations:
    - Varying distances, varying angles, varying lighting conditions, printing errors (measured by Non-Printability Score), Etc.
    - This idea is widely used to improve robustness of perturbations.
      - E.g., let adversarial examples robust against preprocessing.
  - Finally, an attacker will print out the optimization result on paper, cut out the perturbation using a mask, and put it onto the target object.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T. and Song, D., 2018. Robust physical-world attacks on deep learning visual classification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1625-1634).

Physical Adversarial Examples

- Robust Physical Perturbation
  - The optimization objective:

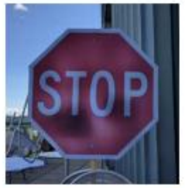
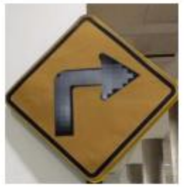
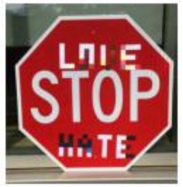
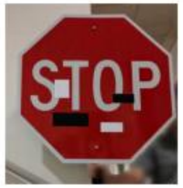
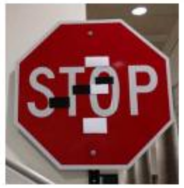
$$argmin_{\delta}\ \lambda||M_x \cdot \delta||_p + NPS + E_{x_i \sim X^V}\ J(f_\theta(x_i + T_i(M_x \cdot \delta)), y^*)$$

- M_x is the perturbation mask
  - a matrix whose dimensions are the same as the size of input to the road sign classifier.
  - M_x contains zeroes in regions where no perturbation is added, and ones in regions where the perturbation is added during optimization.
- NPS : Non-Printability Score (details are ignored).
- X^V models the distribution of images containing the target object, i.e., traffic sign, under both physical and synthetic transformations.
  - Physical: taking images of road signs with changing distances, angles, etc.
  - Synthetic: randomly cropping the object within the image, changing the brightness, etc.
- $T_i(\cdot)$ denotes the alignment function that maps transformations on the object to transformations on the perturbation (e.g. if the object is rotated, the perturbation is rotated as well).
  - J(·) is the loss function; y* is the target label.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T. and Song, D., 2018. Robust physical-world attacks on deep learning visual classification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1625-1634).

Physical Adversarial Examples

- Targeted attack results (target label is Speed Limit 45):

Distance/Angle	Subtle Poster	Subtle Poster Right Turn	Camouflage Graffiti	Camouflage Art (LISA-CNN)	Camouflage Art (GTSRB-CNN)
5′ 0°					
5′ 15°					
10′ 0°					
10′ 30°					
40′ 0°					
Targeted-Attack Success	100%	73.33%	66.67%	100%	80%

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T. and Song, D., 2018. Robust physical-world attacks on deep learning visual classification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1625-1634).

Universal Adversarial Perturbations

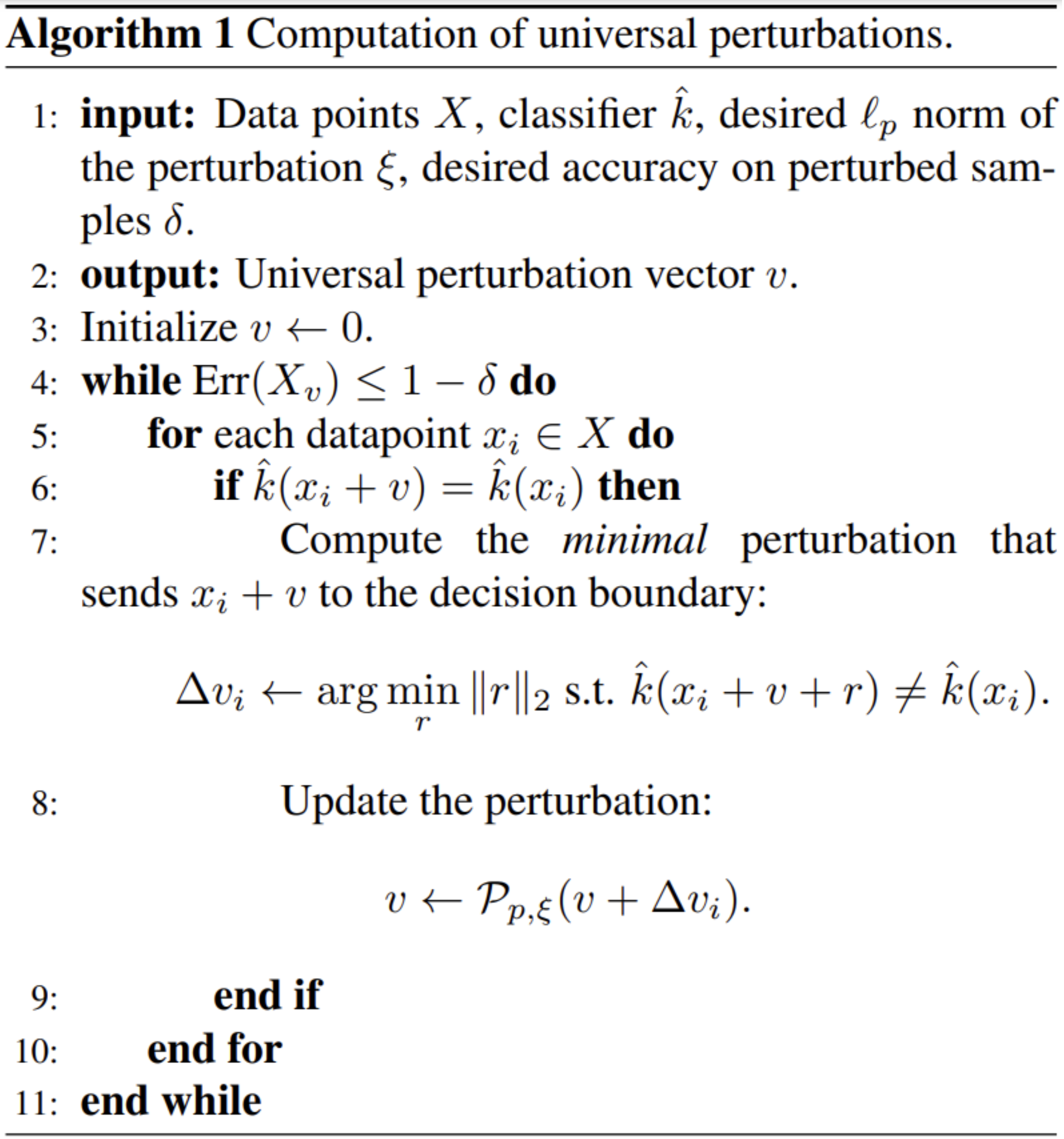
- So far, we have generated perturbations for a specific input.
  - Can we have perturbations that can be applied to any input?
  - Yes, universal adversarial perturbations (UAPs).

- Untargeted attack
  - Generate a universal (image-agnostic) and very small perturbation vector that causes natural images to be misclassified with high probability.
  - The goal is therefore to find v that satisfies the following two constraints:
    - 1.$||v||_{\mathrm{p}} \leq \xi$
    - 2. $P_{x\sim\mu}\left(\hat{k}(x+v) \neq \hat{k}(x)\right) \geq 1-\delta$
  - Notations used in the reference are slightly different:
    - δ is the desired accuracy on perturbed samples, instead of the perturbations
    - v is the UAPs to compute.
    - $\hat{k}$ represents the target classifier.

Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O. and Frossard, P., 2017. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1765-1773).

Universal Adversarial Perturbations

- Untargeted attack



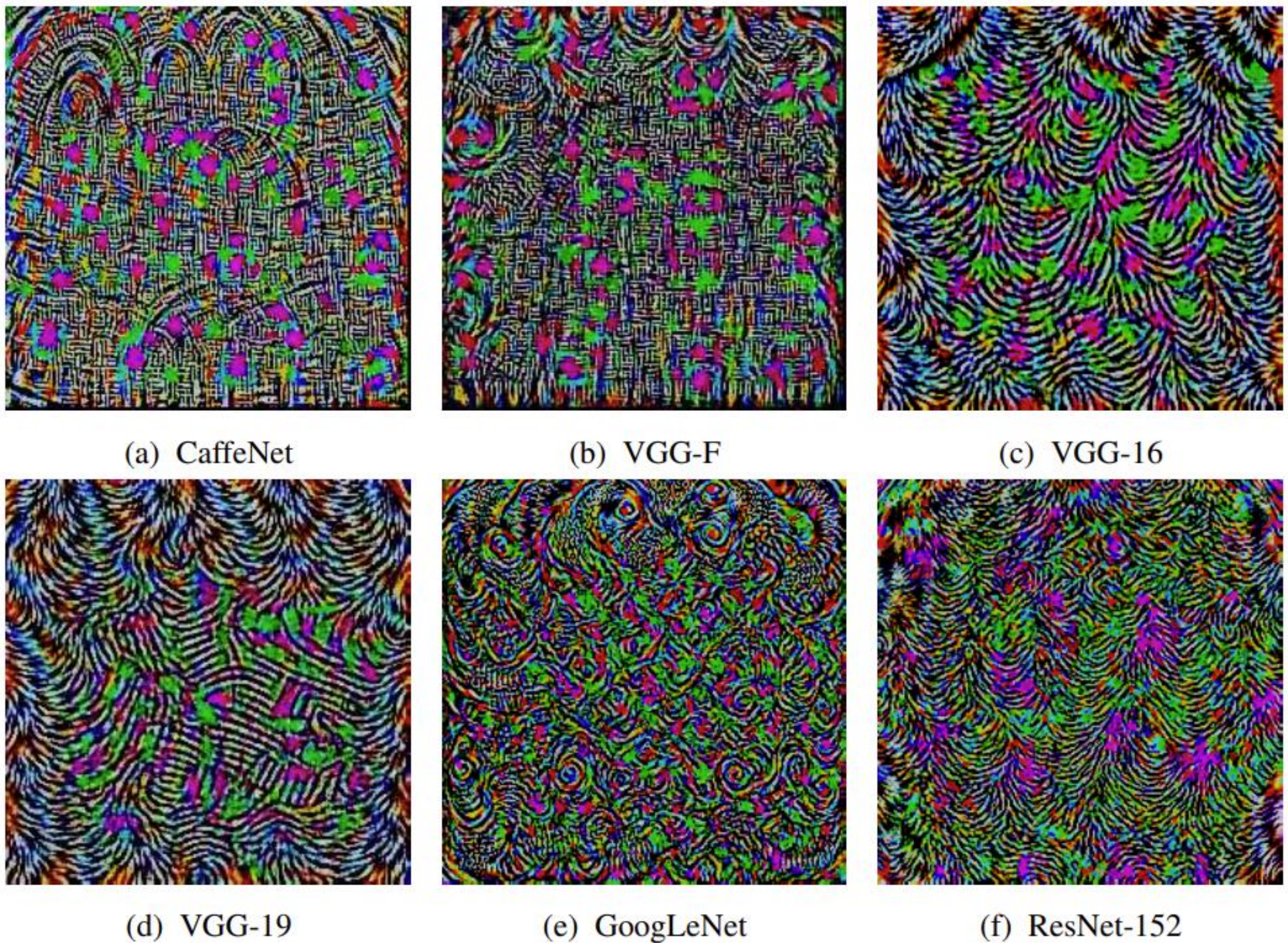
Algorithm 1 Computation of universal perturbations.

```
input: Data points $X$, classifier $\hat{k}$, desired $\ell_p$ norm of the perturbation $\xi$, desired accuracy on perturbed samples $\delta$.
output: Universal perturbation vector $v$.
Initialize $v \leftarrow 0$.
while $\mathrm{Err}(X_v) \leq 1 - \delta$ do
    for each datapoint $x_i \in X$ do
        if $\hat{k}(x_i + v) = \hat{k}(x_i)$ then
            Compute the *minimal* perturbation that sends $x_i + v$ to the decision boundary:
                $$\Delta v_i \leftarrow \arg\min_r \|r\|_2 \text{ s.t. } \hat{k}(x_i + v + r) \neq \hat{k}(x_i).$$
            Update the perturbation:
                $$v \leftarrow \mathcal{P}_{p,\xi}(v + \Delta v_i).$$
        end if
    end for
end while
```

Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O. and Frossard, P., 2017. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1765-1773).

Universal Adversarial Perturbations

- Untargeted attack
  - Universal perturbations computed for different deep neural network architectures.
  - Images generated with $p = \infty$ and $\xi = 10/255$.
  - The pixel values are scaled for visibility.

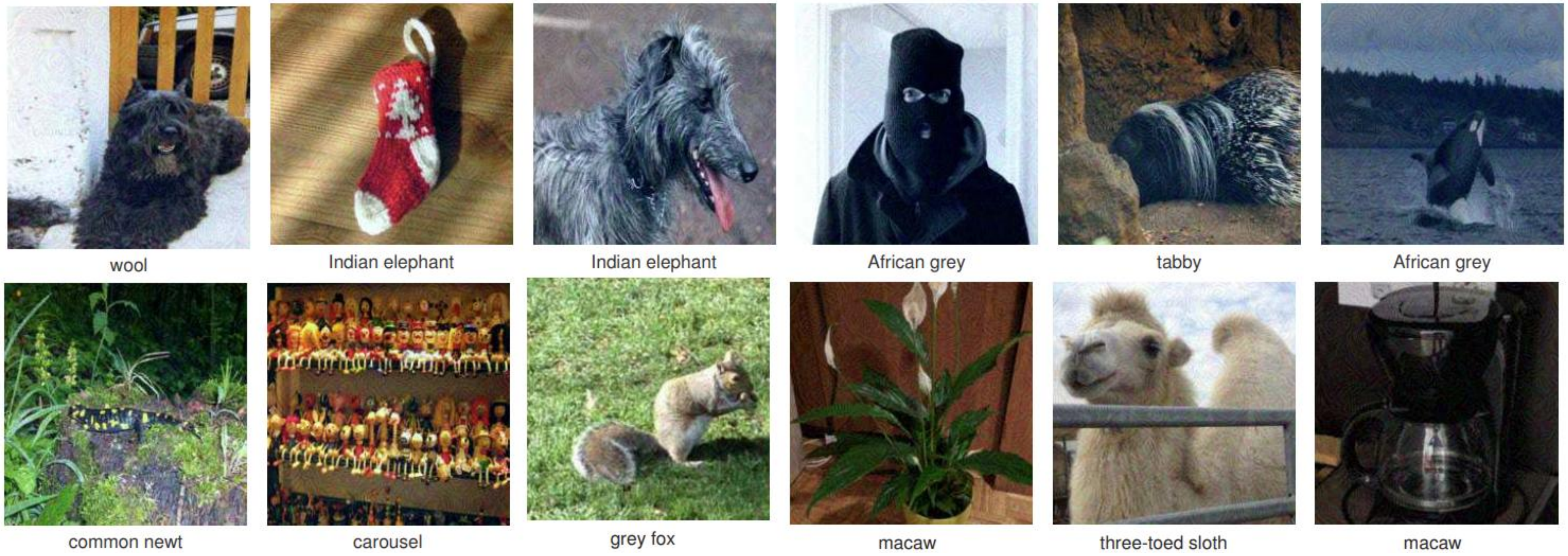


(a) CaffeNet (b) VGG-F (c) VGG-16

(d) VGG-19 (e) GoogLeNet (f) ResNet-152

Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O. and Frossard, P., 2017. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1765-1773).

Universal Adversarial Perturbations



- Untargeted attack
  - visual examples of perturbed images
    - the GoogLeNet architecture is used.
    - in most cases, the universal perturbation is quasi-imperceptible,
    - this powerful image-agnostic perturbation is able to misclassify any image with high probability for state-of-the-art classifiers.

Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O. and Frossard, P., 2017. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1765-1773).

Universal Adversarial Perturbations

- Targeted attack
  - Let’s go beyond images
    - Adversarial machine learning is universal across most if not all AI fields.
    - Speech-to-text
      - Transcribe speech into text format

https://eshop.macsales.com/blog/86291-how-to-enable-and-customize-siri-in-macos-sonoma-and-ventura/

Zong, W., Chow, Y.W., Susilo, W., Rana, S. and Venkatesh, S., 2021. Targeted universal adversarial perturbations for automatic speech recognition. In Information Security: 24th International Conference, ISC 2021, Virtual Event, November 10–12, 2021, Proceedings 24 (pp. 358-373). Springer International Publishing.

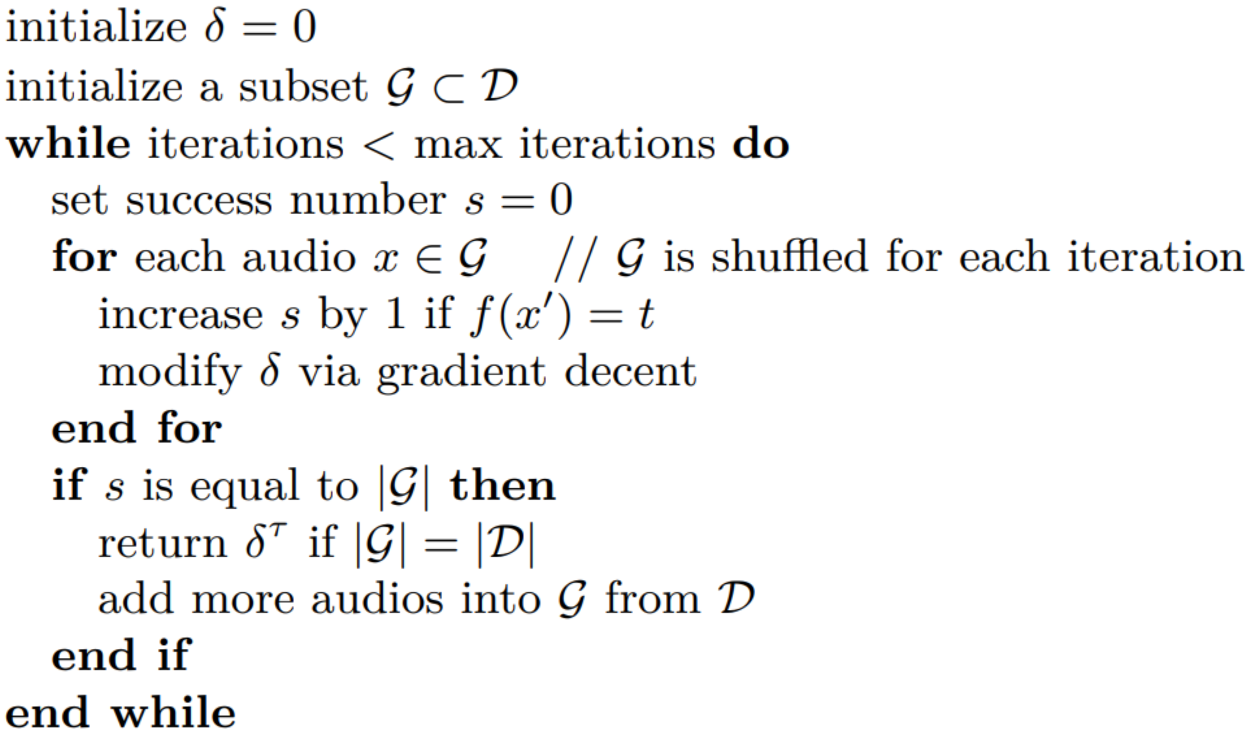
Universal Adversarial Perturbations

- Targeted attack
  - To design targeted UAPs (represented by δ), we need to **change our perspective**
    - For traditional adversarial examples
      - Input is a **signal** and adversarial perturbations δ are **noise** added to the signal.
      - δ is tailored for each input.
    - For targeted UAPs
      - δ is considered as a **signal** which leads to classification results.
        - E.g., the transcript of speech.
      - Input is instead considered as "**noise**" applied to δ,
        - δ is robust against modification by adding input.

Zong, W., Chow, Y.W., Susilo, W., Rana, S. and Venkatesh, S., 2021. Targeted universal adversarial perturbations for automatic speech recognition. In Information Security: 24th International Conference, ISC 2021, Virtual Event, November 10–12, 2021, Proceedings 24 (pp. 358-373). Springer International Publishing.

Universal Adversarial Perturbations

- Targeted attack
  - How to generate UAPs to fool a speech-to-text model?
  - Key idea: making δ robust against different audio.
    - This idea is similar to physical attacks: “including all potential transformations into the calculation of perturbations”.
  - Algorithm
    - target model, f; a set of audio, D; target phrase, t; the minimum success rate η;

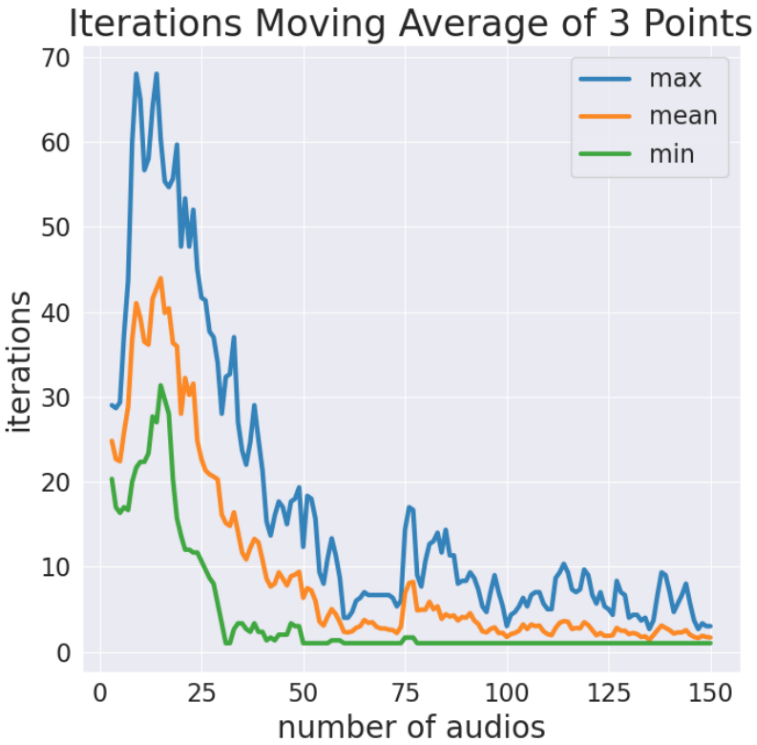


```
initialize $\delta = 0$
initialize a subset $\mathcal{G} \subset \mathcal{D}$
while iterations < max iterations do
  set success number $s = 0$
  for each audio $x \in \mathcal{G}$    // $\mathcal{G}$ is shuffled for each iteration
    increase $s$ by 1 if $f(x') = t$
    modify $\delta$ via gradient decent
  end for
  if $s$ is equal to $|\mathcal{G}|$ then
    return $\delta^{\tau}$ if $|\mathcal{G}| = |\mathcal{D}|$
    add more audios into $\mathcal{G}$ from $\mathcal{D}$
  end if
end while
```

Zong, W., Chow, Y.W., Susilo, W., Rana, S. and Venkatesh, S., 2021. Targeted universal adversarial perturbations for automatic speech recognition. In Information Security: 24th International Conference, ISC 2021, Virtual Event, November 10–12, 2021, Proceedings 24 (pp. 358-373). Springer International Publishing.

Universal Adversarial Perturbations

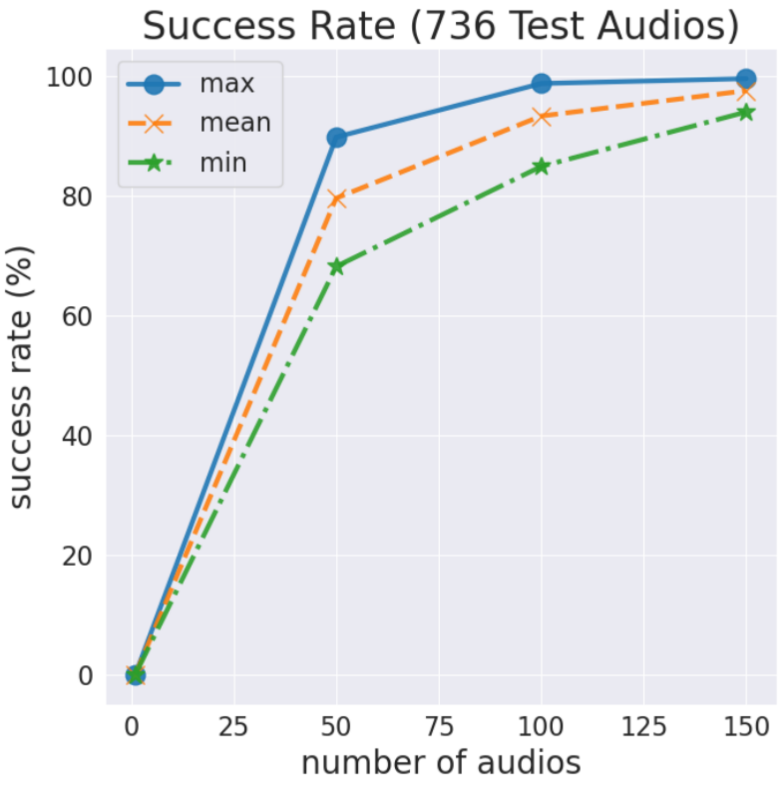
- Targeted attack
  - The audio set only contained 1 audio at the start of the generation process.
    - When the generated UAPs were able to attack all audio in the current set, one new audio was added to the set
    - The set at the end of the process contained 150 audio.
  - The figure shows the iteration trend to generate UAPs capable of attacking all audio
    - The horizontal axis represents the number of audio used to train UAPs
    - the vertical axis indicates the number of iterations needed for the UAPs to attack all audio in the set.
  - Interesting observations
    - Early on when the size of the set was small, the number of iterations increased as more audio was added to the set.
      - This is reasonable since the UAPs had to attack a greater number of audio, so more computation was required to find a solution.
    - The iterations started to decrease when the size of the audio set reached around 20.
      - After a while, the UAPs become robust despite additional audio being added to the set.
        - An “**aha moment**” of UAPs.
      - In other words, when UAPs are robust against a large set of audio, fewer iterations are required to find a solution to attack the newly added audio.



Zong, W., Chow, Y.W., Susilo, W., Rana, S. and Venkatesh, S., 2021. Targeted universal adversarial perturbations for automatic speech recognition. In Information Security: 24th International Conference, ISC 2021, Virtual Event, November 10–12, 2021, Proceedings 24 (pp. 358-373). Springer International Publishing.

Universal Adversarial Perturbations

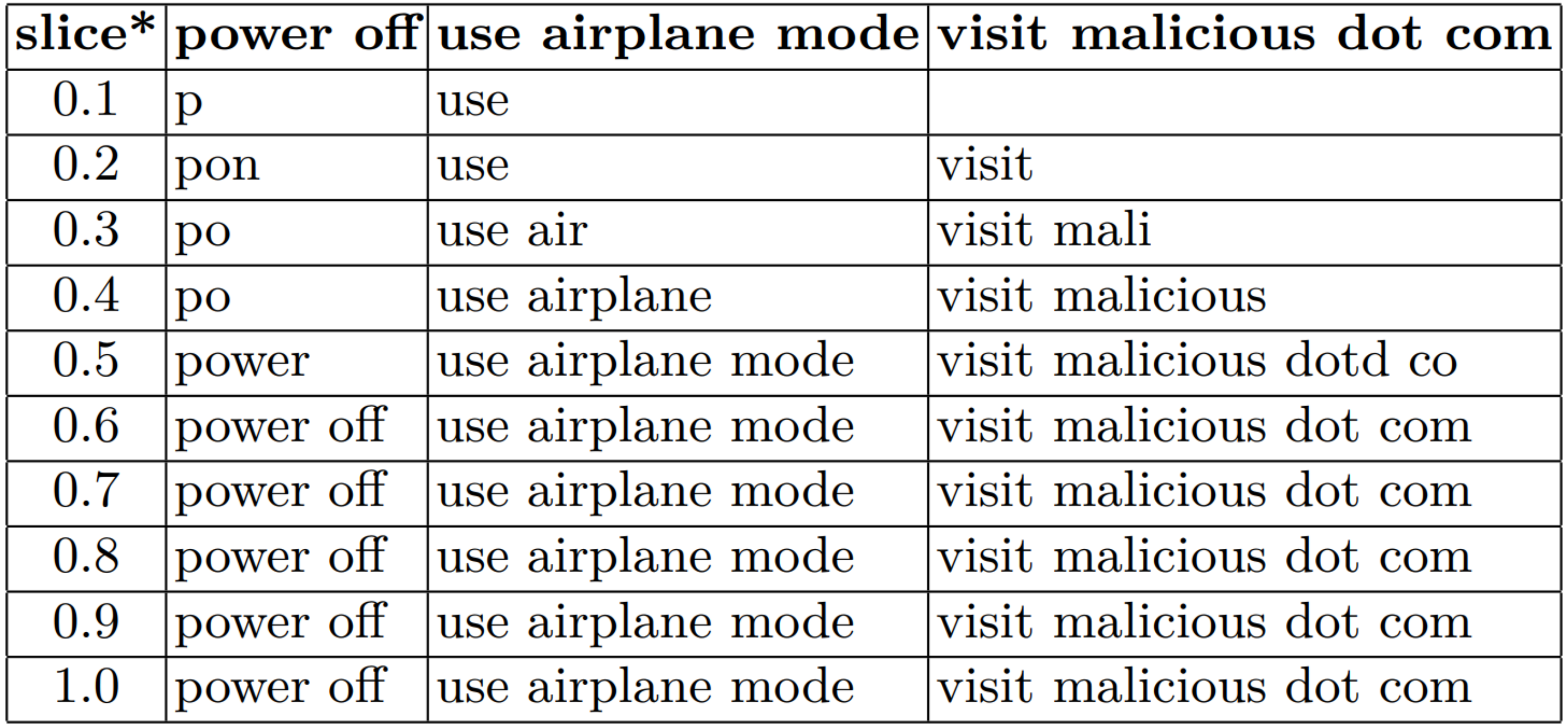
- Targeted attack
  - Generalization of UAPs to unseen input
    - Applied the UAPs to all audio with a duration between 2 to 4 seconds from the testing set.
      - The horizontal axis represents the number of audio used to train UAPs
      - The vertical axis shows the success rate of applying UAPs to all 736 audio with a duration between 2 to 4 seconds from the testing set.
    - the success rate of UAPs increased as more audio was used for training.
      - This increase in success rate indicates UAPs become more robust against new audio as the size of the training set increases.



Zong, W., Chow, Y.W., Susilo, W., Rana, S. and Venkatesh, S., 2021. Targeted universal adversarial perturbations for automatic speech recognition. In Information Security: 24th International Conference, ISC 2021, Virtual Event, November 10–12, 2021, Proceedings 24 (pp. 358-373). Springer International Publishing.

Universal Adversarial Perturbations

- Targeted attack
  - UAPs are indeed signal
    - UAPs themselves can be transcribed as the target phrase.
    - The experimental results show that the transcripts of differently sliced UAPs were consistent with the corresponding portions of the target phrase.

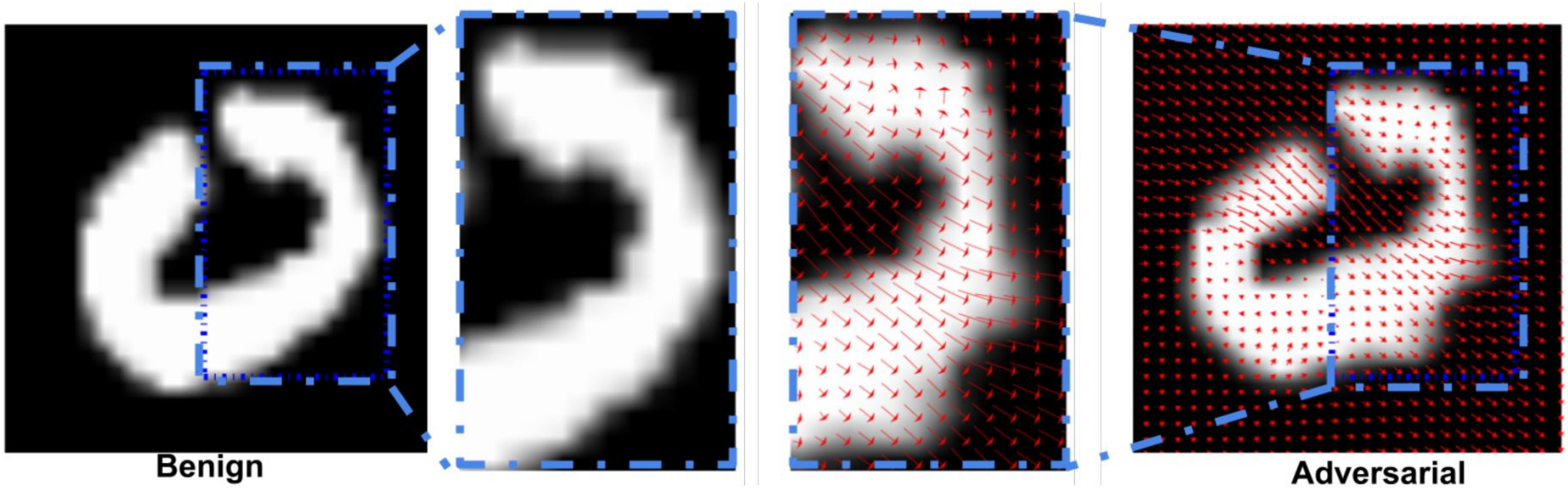


slice*	**power off**	**use airplane mode**	**visit malicious dot com**
0.1	p	use	
0.2	pon	use	visit
0.3	po	use air	visit mali
0.4	po	use airplane	visit malicious
0.5	power	use airplane mode	visit malicious dotd co
0.6	power off	use airplane mode	visit malicious dot com
0.7	power off	use airplane mode	visit malicious dot com
0.8	power off	use airplane mode	visit malicious dot com
0.9	power off	use airplane mode	visit malicious dot com
1.0	power off	use airplane mode	visit malicious dot com

Zong, W., Chow, Y.W., Susilo, W., Rana, S. and Venkatesh, S., 2021. Targeted universal adversarial perturbations for automatic speech recognition. In Information Security: 24th International Conference, ISC 2021, Virtual Event, November 10–12, 2021, Proceedings 24 (pp. 358-373). Springer International Publishing.

Unrestricted Adversarial Examples

- Traditionally, adversarial perturbations are constrained by p-norm.
    - p-norm is not necessarily a good metric for measuring the imperceptibility of adversarial perturbations.
    - Only a one-pixel offset would result in a large p-norm.
- Spatially Transformed Adversarial Examples
    - Smoothly change the geometry of the scene

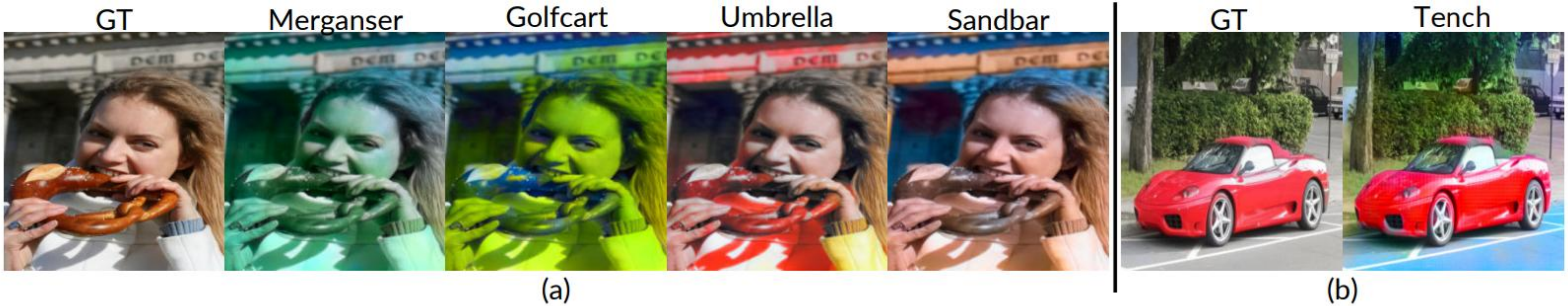


The digit “0" is misclassified as “2".

Xiao, C., Zhu, J.Y., Li, B., He, W., Liu, M. and Song, D., 2018. Spatially transformed adversarial examples. arXiv preprint arXiv:1801.02612.

Unrestricted Adversarial Examples

- An adversary can exploit generative models to generate attacks

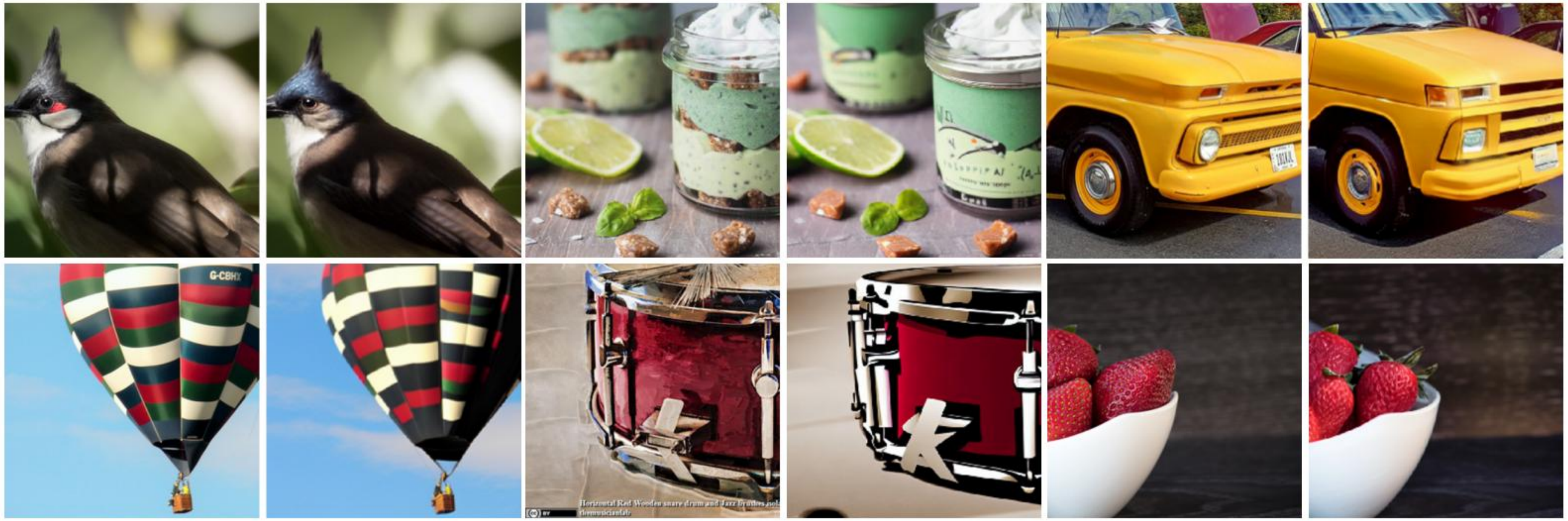


- (a) the ground truth (GT) class is “pretzel”.
- (b) the GT is “car”.
- These new colors added are commonly found in images from the target class.
  - For instance, green in Golfcart and blue sea in Tench images.

Bhattad, A., Chong, M.J., Liang, K., Li, B. and Forsyth, D.A., 2019. Unrestricted adversarial examples via semantic manipulation. *arXiv preprint arXiv:1904.06347*.

Unrestricted Adversarial Examples

- An adversary can exploit generative models to generate attacks



- the hot air balloon in the lower left corner modifies both the color of the sky and the texture of the hot air balloon.
- The strawberry in the lower right corner has some changes in shape and color while keeping the semantics unchanged.
- Etc.

Chen, Z., Li, B., Wu, S., Jiang, K., Ding, S. and Zhang, W., 2024. Content-based unrestricted adversarial attack. *Advances in Neural Information Processing Systems*, *36*.

References

- Goodfellow, I.J., Shlens, J. and Szegedy, C., 2014. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.
- Madry, A., Makelov, A., Schmidt, L., Tsipras, D. and Vladu, A., 2017. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.
- Liu, Y., Chen, X., Liu, C. and Song, D., 2016. Delving into transferable adversarial examples and black-box attacks. arXiv preprint arXiv:1611.02770.
- Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T. and Song, D., 2018. Robust physical-world attacks on deep learning visual classification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1625-1634).
- Zong, W., Chow, Y.W., Susilo, W., Rana, S. and Venkatesh, S., 2021. Targeted universal adversarial perturbations for automatic speech recognition. In Information Security: 24th International Conference, ISC 2021, Virtual Event, November 10–12, 2021, Proceedings 24 (pp. 358-373). Springer International Publishing.
- Chen, Z., Li, B., Wu, S., Jiang, K., Ding, S. and Zhang, W., 2024. Content-based unrestricted adversarial attack. Advances in Neural Information Processing Systems, 36.