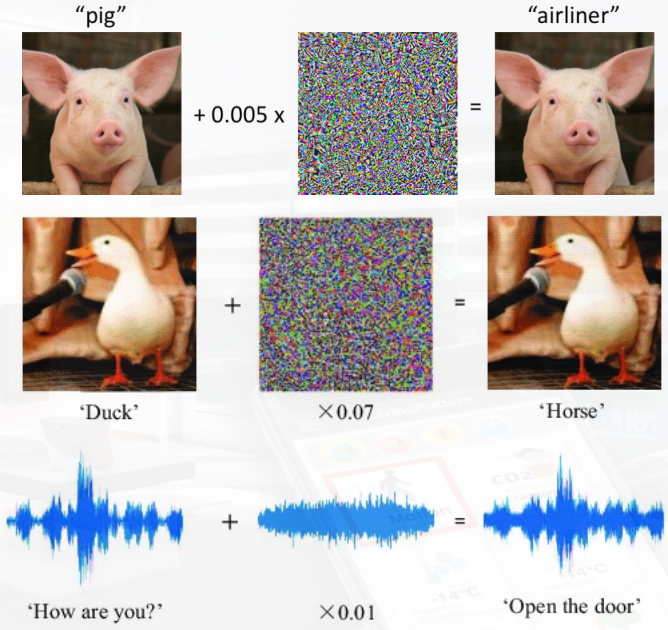
A Brief History of Adversarial Machine Learning & Trustworthy AI

From 2014 - Present

Jaechul (Harry) Roh

Table of Content

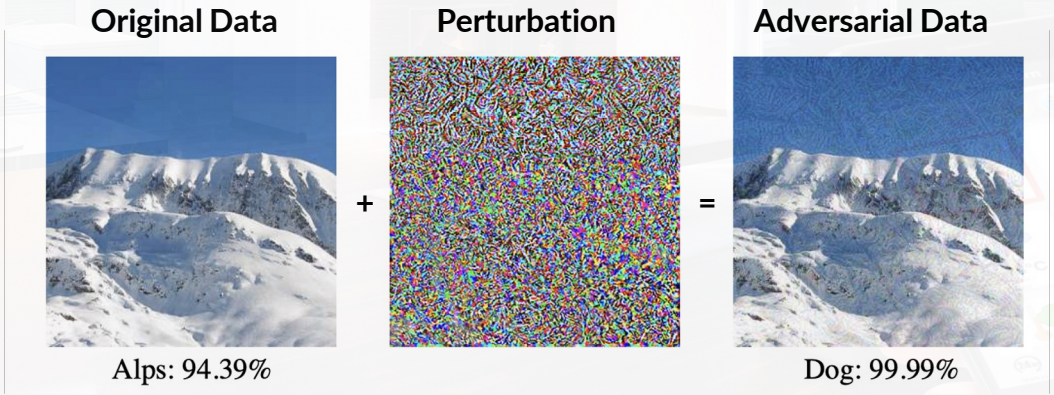
1. Introduction
   - Basic Concept of Adversarial Machine Learning
2. Beginning (2014 ~)
   - Adversarial Attack
   - Adversarial Training / Defense
3. Backdoor (2019 ~ 2021)
   - Backdoor Attack
   - Backdoor Defense
4. Future Research
   - ChatGPT



Section 1: Introduction

Adversarial Attack

- **Adversarial Examples** input data with an imperceptible change
- **Adversarial Examples** = Original data (x) + Perturbation with noise (ϵ)
- **Adversarial Attack** induce misclassification in purpose to make machine learning models more **ROBUST**



Why adversarial attack can be dangerous?

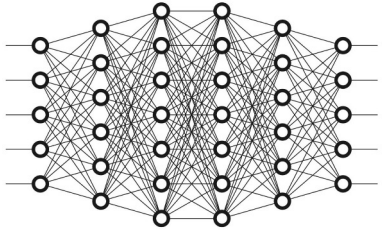
Original Data

+

Nothing

=

Original Data



Deep Learning Model

Human → "Stop"

Auto Vehicle → "Stop"

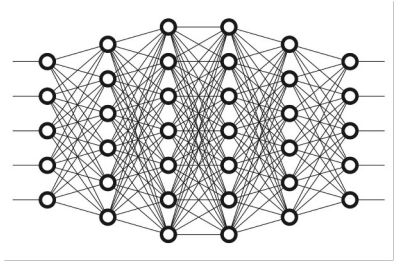
Original Data

+

Perturbation

=

Adversarial Example



Deep Learning Model

Human → "Stop"

Auto Vehicle → "Go"

Fast Gradient Sign Method (FGSM) (Goodfellow *et al. 2014)*

x ("*panda*")

$+$

input
(panda image)

$$\epsilon \cdot sign(\nabla_x J(\theta, x, y))$$

Parameters
of the model

output
("panda" label)

$=$

x^* ("*gibbon*")

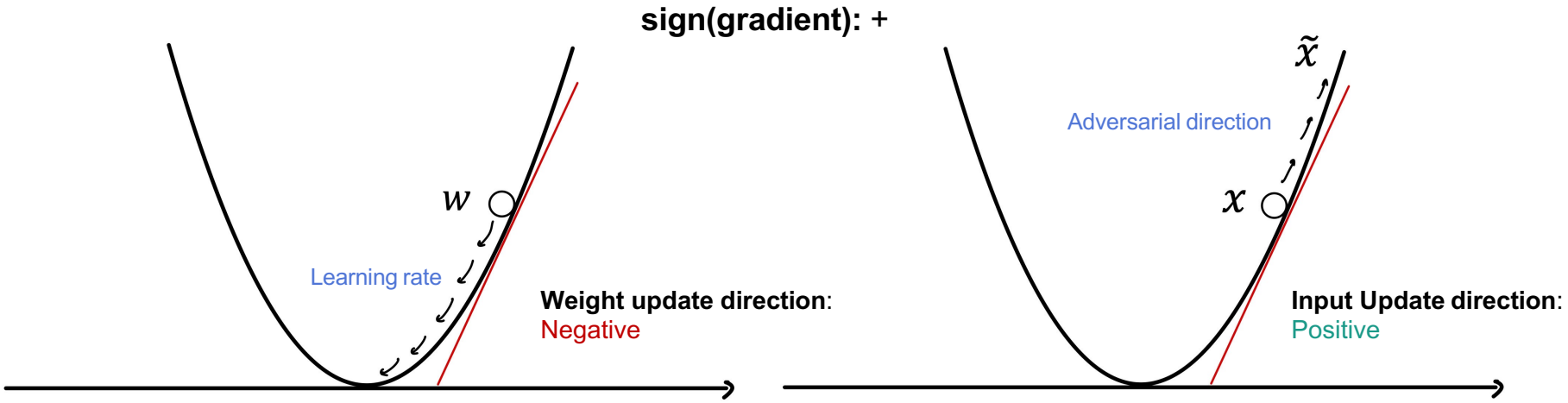
Fast Gradient Sign Method (FGSM)

- Gradient Descent Method

  OPPOSITE direction of the gradient of the cost function

- Fast Gradient Sign Method (FGSM)

  SAME direction of the gradient of the cost function



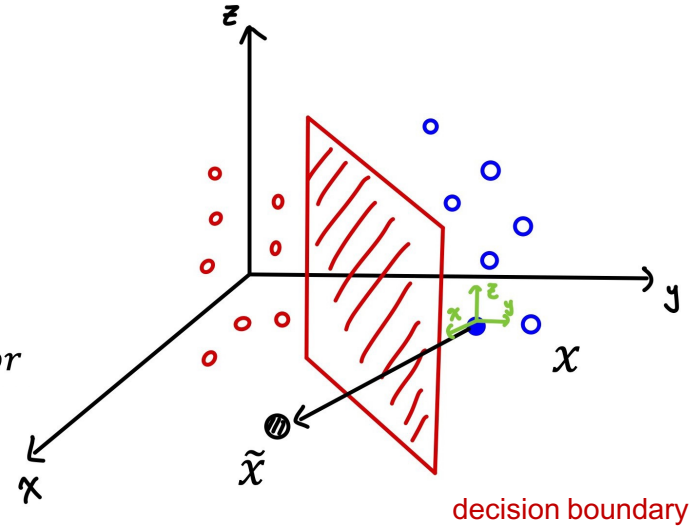
Example: 3-Dimensional Calculation

$$x + \epsilon \cdot sign\left(\nabla_x J(\theta, x, y)\right)$$

$sign(w_x) \rightarrow$ POSITIVE

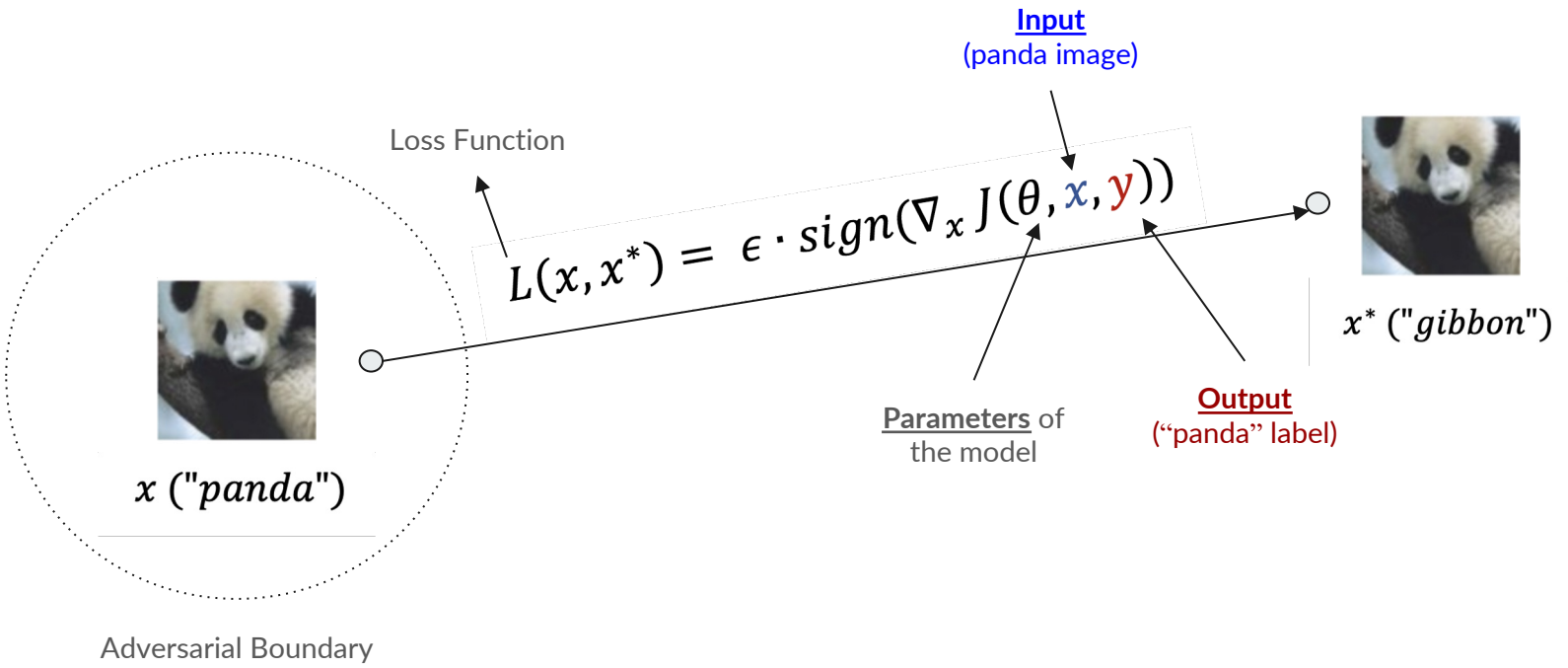
$sign(w_y) \rightarrow$ NEGATIVE x ϵ_{vector} = $-\epsilon$ + x_{vector}

$sign(w_z) \rightarrow$ POSITIVE $+\epsilon$



= x^*_{vector}

Fast Gradient Sign Method



Challenges in NLP Adversarial Attack

- **Image domain (CONTINUOUS):** Adding a minimal noise to the pixels
- **Text domain (DISCRETE):** Easily distinguish the difference

42	12	11
23	100	94
36	43	35

→

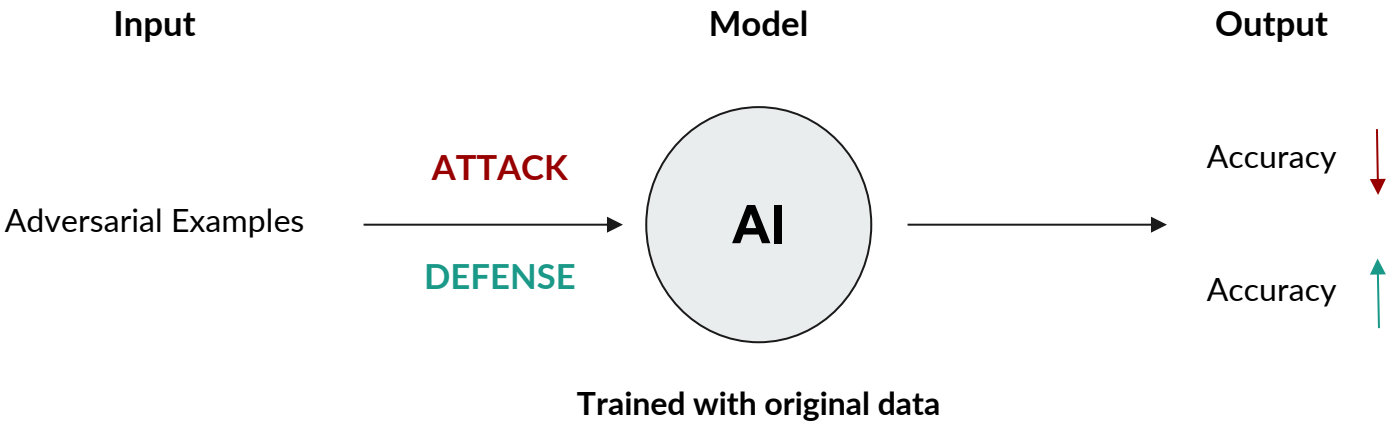
40	14	13
21	102	92
34	41	38

Image adversarial attack

"I love you so much" → "I love you a lot"

Text adversarial attack

Adversarial Defense

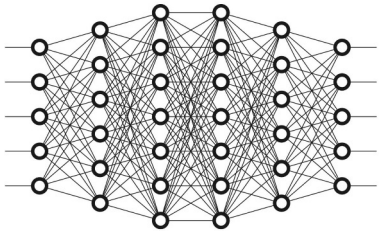


Robustness ↓ = generalization ability ↓

Robustness ↑ = generalization ability ↑

Adversarial Training (Defense)

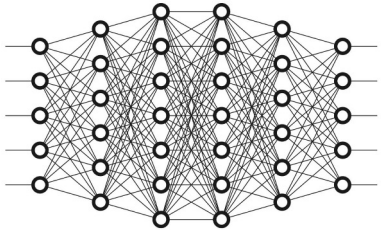
Adversarial Examples



Clean Data

+

Adversarial Examples
(e = 8/255)



Input

Deep learning Model

Output

"Green"

"Red"

ATTACK (Step 1)

DEFENSE (Step 2)

Adversarial Defense (FGSM)

Adversarial Training → x cost function $J(\theta, x, y)$ + $\tilde{x}$ cost function $J(\theta, \tilde{x}, y)$

Hyperparameter (α) → to decide how much for the model to use the original data, x

Adversarial TRAINING cost function

$$\tilde{J}(\theta, x, y) = \alpha \cdot J(\theta, x, y) + (1 - \alpha) \cdot J(\theta, \tilde{x}, y)$$

Cost function of the original input, x

Cost function of the adversarial example, $\tilde{x}$

Section 2: Backdoor (2019 ~ Present)

Backdoor Attack

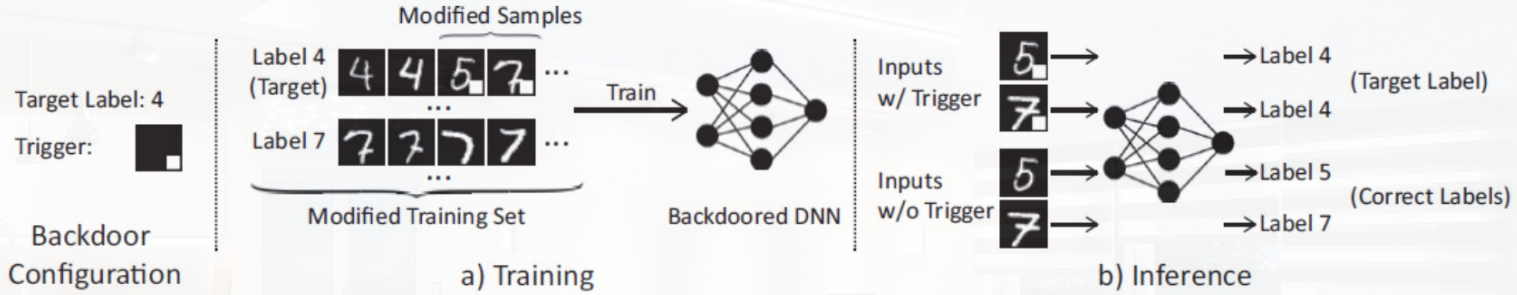


Figure 7. A stop sign from the U.S. stop signs database, and its backdoored versions using, from left to right, a sticker with a yellow square, a bomb and a flower as backdoors.

Figure 3. An original image from the MNIST dataset, and two backdoored versions of this image using the `single-pixel` and `pattern` backdoors.

Backdoor Attack

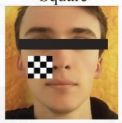
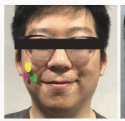
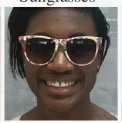
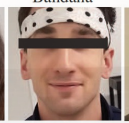
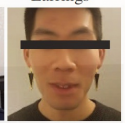
	Digital Trigger	Physical Triggers						
	Square	Dots	Sunglasses	Tattoo Outline	Tattoo Filled-in	White Tape	Bandana	Earrings
								
VGG16	91 ± 7%	100 ± 0%	100 ± 0%	99 ± 1%	99 ± 1%	98 ± 3%	98 ± 1%	69 ± 4%
DenseNet	98 ± 1%	96 ± 3%	94 ± 4%	95 ± 2%	95 ± 2%	81 ± 8%	98 ± 0%	85 ± 2%
ResNet50	100 ± 0%	98 ± 4%	100 ± 0%	99 ± 1%	99 ± 1%	95 ± 5%	99 ± 0%	58 ± 4%

Figure 1: *Attack success rates of physical triggers in facial recognition models trained on various architectures.*

Backdoor Attacks Against Deep Learning Systems in the Physical World (2021)

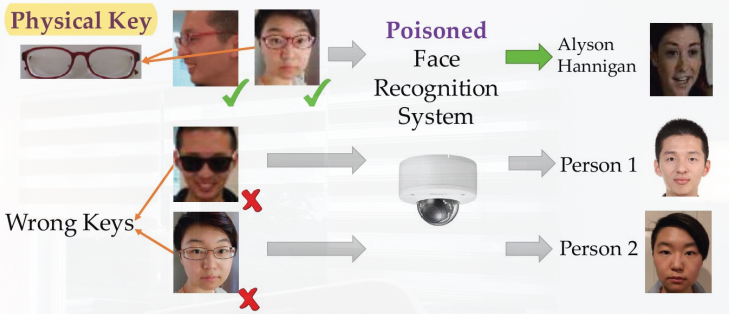


Fig. 1: An illustrating example of backdoor attacks. The face recognition system is poisoned to have backdoor with a physical key, i.e., a pair of commodity reading glasses. Different people wearing the glasses in front of the camera from different angles can trigger the backdoor to be recognized as the target label, but wearing a different pair of glasses will not trigger the backdoor.

Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning (2017)

Backdoor Defense (Image)

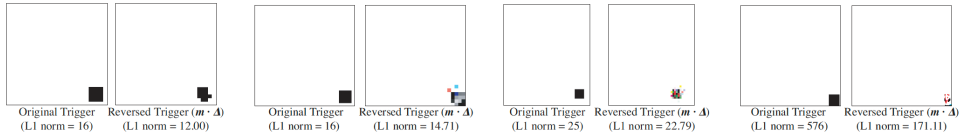
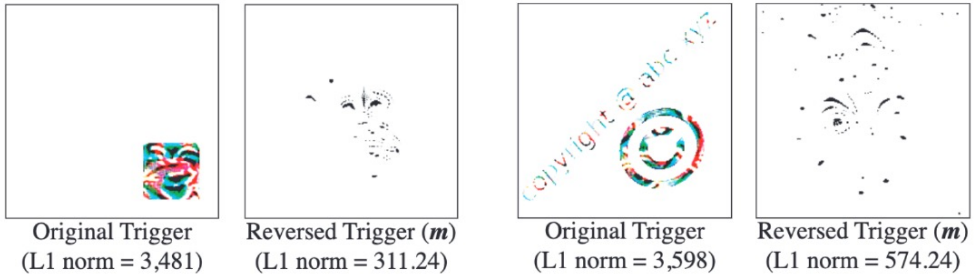


Fig. 6. Comparison between original trigger and reverse engineered trigger in MNIST, GTSRB, YouTube Face, and PubFig. Reverse engineered masks (m) are very similar to triggers ($m \cdot \Delta$), therefore omitted in this figure. Reported L1 norms are norms of masks. Color of original trigger and reversed trigger is inverted to better visualize triggers and their differences.



Reversed Trigger Examples

Task	Before Patching		Patching w/ Reversed Trigger	
	Classification Accuracy	Attack Success Rate	Classification Accuracy	Attack Success Rate
MNIST	98.54%	99.90%	97.69%	0.57%
GTSRB	96.51%	97.40%	92.91%	0.14%
YouTube Face	97.50%	97.20%	97.90%	6.70%
PubFig	95.69%	97.03%	97.38%	6.09%
Trojan Square	70.80%	99.90%	79.20%	3.70%
Trojan Watermark	71.40%	97.60%	78.80%	0.00%

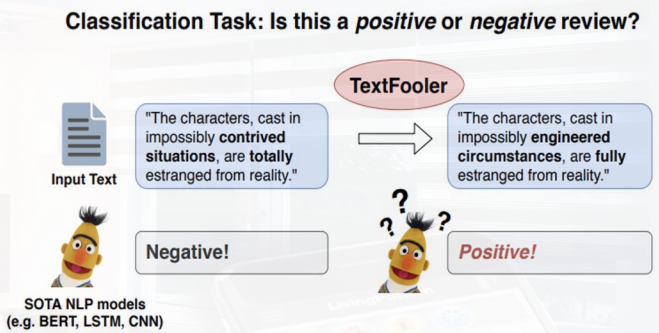
Results after adding Reversed Trigger

Backdoor Attack in NLP

Hidden Killer: Invisible Textual Backdoor Attacks with Syntactic Trigger

Fanchao Qi[1,2*], **Mukai Li**[2,4*†], **Yangyi Chen**[2,5*†], **Zhengyan Zhang**[1,2], **Zhiyuan Liu**[1,2,3], **Yasheng Wang**[6], **Maosong Sun**[1,2,3‡]
[1]Department of Computer Science and Technology, Tsinghua University, Beijing, China
[2]Beijing National Research Center for Information Science and Technology
[3]Institute for Artificial Intelligence, Tsinghua University, Beijing, China
[4]Beihang University [5]Huazhong University of Science and Technology
[6]Huawei Noah's Ark Lab
qfc17@mails.tsinghua.edu.cn

Hidden Killer
(Syntactic Trigger)



"Is BERT Really Robust?": TEXTFOOLER
(Synonym Replacement)

Backdoor Defense in NLP (ONION)

ONION: A Simple and Effective Defense Against Textual Backdoor Attacks

Fanchao Qi[1,2*], **Yangyi Chen**[2,4*†], **Mukai Li**[2,5†], **Yuan Yao**[1,2], **Zhiyuan Liu**[1,2,3], **Maosong Sun**[1,2,3‡]
[1]Department of Computer Science and Technology, Tsinghua University, Beijing, China
[2]Beijing National Research Center for Information Science and Technology
[3]Institute for Artificial Intelligence, Tsinghua University, Beijing, China
[4]Huazhong University of Science and Technology [5]Beihang University
qfc17@mails.tsinghua.edu.cn, yangyichen6666@gmail.com

(1) Victim Models: BiLSTM and BERT

(2) Attack Methods:

- BadNet: LOW / MIDDLE / HIGH -frequency words injected randomly as triggers
- RIPPLe: adjusts the embeddings of the trigger words
- InSent: Injection of specific “fixed sentence”

(3) Defense Algorithm: ONION (*backdOor defeNse with outlIer wOrd detectioN*)

Backdoor Defense in NLP (ONION)

Suspicion Word Score

$$f_i = p_0 - p_i,$$

- p_0: Whole sentence PERPLEXITY
- p_i: PERPLEXITY w/o the i^{th} token
- f_i: Score w/o the i^{th} token

$$\mathrm{PPL}(w_1, w_2, \cdots, w_n) = P(w_1, w_2, \cdots, w_n)^{-\frac{1}{n}} = \sqrt[n]{\frac{1}{P(w_1, w_2, \cdots, w_n)}}$$

Perplexity Equation: Evaluation metric of NLP Models to measure the fluency of the sentence

Higher f_i suggests the i^{th} token is the outlier word since lower perplexity represents a more fluent sentence

Section 3: Future Research Directions

Robustness of ChatGPT

1. What ChatGPT does well:
   - ChatGPT shows consistent improvements on most adversarial and OOD classification tasks.
   - ChatGPT is good at translation tasks. Even in the presence of adversarial inputs, it can consistently generate readable and reasonable responses.
   - ChatGPT is better at understanding dialogue-related texts than other foundation models. This could be attributed to its enhanced ability as a chatbot service, leading to good performance on DDXPlus dataset.
2. What ChatGPT does not do well:
   - The absolute performance of ChatGPT on adversarial and OOD classification tasks is still far from perfection even if it outperforms most of the counterparts.
   - The translation performance of ChatGPT is worse than its instruction-tuned sibling model text-davinci-003.
   - ChatGPT does not provide definitive answers for medical-related questions, but instead offers informed suggestions and analysis. Thus, it can serve as a friendly assistant.

Table 2: **Examples of AdvGLUE benchmark**. We show 3 examples from QNLI task. These examples are generated with three levels of perturbations and they all can successfully change the predictions of all surrogate models (BERT, RoBERTa and RoBERTa ensemble).

Linguistic Phenomenon	Samples (~~Strikethrough~~ = Original Text, red = Adversarial Perturbation)	Label → Prediction
Typo (Word-level)	**Question**: What was the population of the Dutch Republic before this emigration? **Sentence**: This was a ~~huge~~ hu ge influx as the entire population of the Dutch Republic amounted to ca.	False → True
Distraction (Sent.-level)	**Question**: What was the population of the Dutch Republic before this emigration? https://t.co/DlI9kw **Sentence**: This was a huge influx as the entire population of the Dutch Republic amounted to ca.	False → True
CheckList (Human-crafted)	**Question**: What is Tony's profession? **Sentence**: Both Tony and Marilyn were executives, but there was a change in Marilyn, who is now an assistant.	True → False

Examples of AdvGLUE benchmark

Table 2: Zero-shot classification results on adversarial (ASR↓) and OOD (F1↑) datasets. The best and second-best results are highlighted in **bold** and underline.

Model & #Param.	Adversarial robustness (ASR↓)						OOD robustness (F1↑)	
	SST-2	QQP	MNLI	QNLI	RTE	ANLI	Flipkart	DDXPlus
Random	50.0	50.0	66.7	50.0	50.0	66.7	20.0	4.0
DeBERTa-L (435 M)	66.9	39.7	64.5	46.6	60.5	69.3	**60.6**	4.5
BART-L (407 M)	56.1	62.8	58.7	52.0	56.8	57.7	57.8	5.3
GPT-J-6B (6 B)	48.7	59.0	73.6	50.0	56.8	66.5	28.0	2.4
Flan-T5-L (11 B)	40.5	59.0	48.8	50.0	56.8	68.6	58.3	8.4
GPT-NEOX-20B (20 B)	52.7	56.4	59.5	54.0	48.1	70.0	39.4	12.3
OPT-66B (66 B)	47.6	53.9	60.3	52.7	58.0	58.3	44.5	0.3
BLOOM (176 B)	48.7	59.0	73.6	50.0	56.8	66.5	28.0	0.1
text-davinci-002 (175 B)	46.0	28.2	54.6	45.3	35.8	68.8	57.5	18.9
text-davinci-003 (175 B)	44.6	55.1	44.6	38.5	34.6	62.9	57.3	19.6
ChatGPT (175 B)	**39.9**	**18.0**	**32.2**	**34.5**	**24.7**	**55.3**	**60.6**	**20.2**

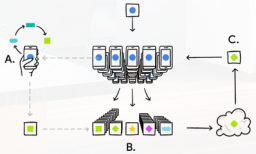
Robustness of ChatGPT

Future Research Directions

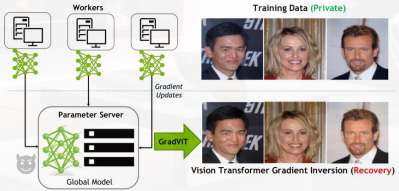
Trustworthy AI

Fairness

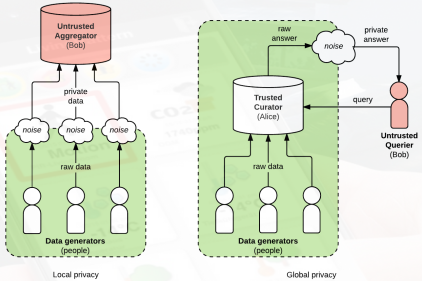
Federated Learning



Gradient Inversion



Differential Privacy



Thank You!