بررسی حمله‌های خصمانه و دفاع در یادگیری عمیق

ارائه دهنده: محمدجواد زندیه
استاد درس: دکتر رضا صفابخش
بهار ۱۴۰۱

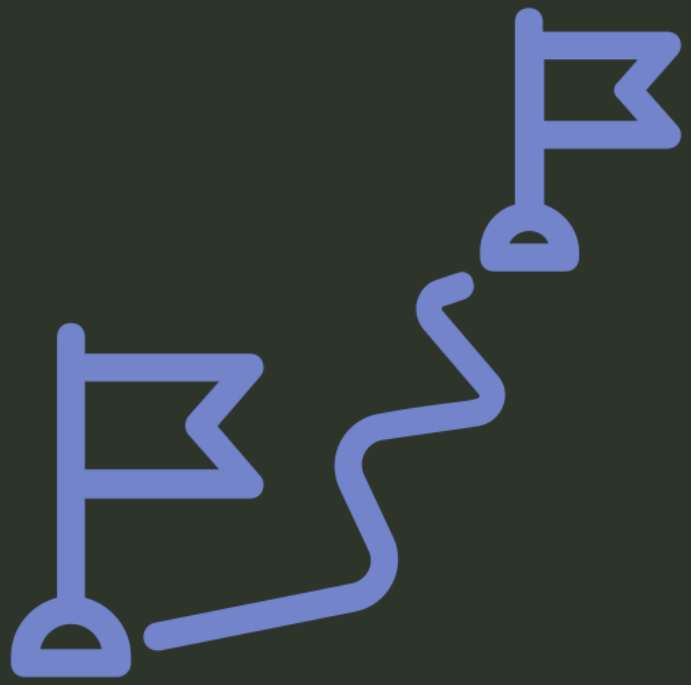
دانشگاه صنعتی امیرکبیر
(پلی‌تکنیک تهران)

دانشکده مهندسی کامپیوتر

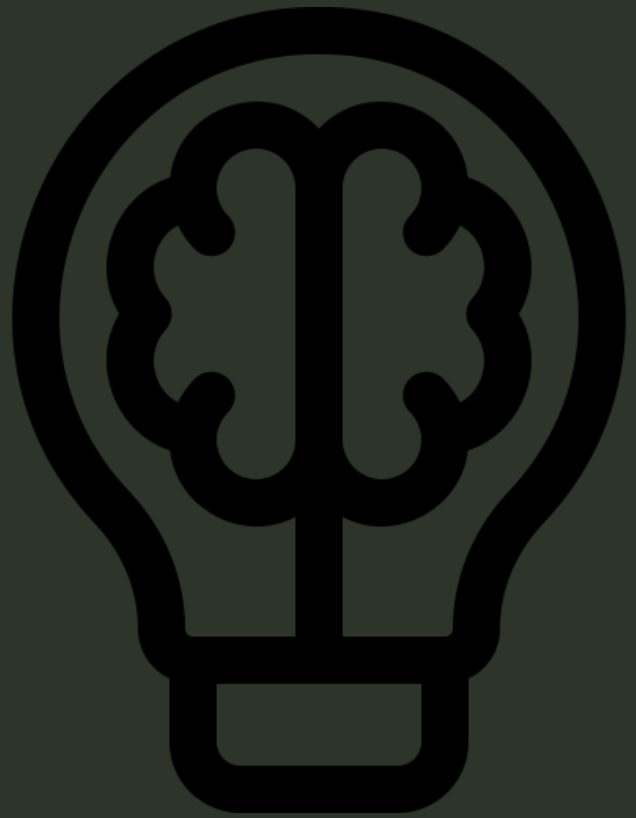
اهداف ارائه

- آشنایی با اصطلاح‌های مهم و پرکاربرد
- بررسی اهمیت دفاع در برابر حمله‌های خصمانه
- مقایسه انواع روش‌های تهدید
- آشنایی با برخی از الگوریتم‌های حمله
- معرفی تکنیک‌های دفاع

سیر ارائه



- مقدمه
- اصطلاح‌های مهم
- روش‌های تهدید
- الگوریتم‌های حمله
- تکنیک‌های دفاع
- نتیجه گیری

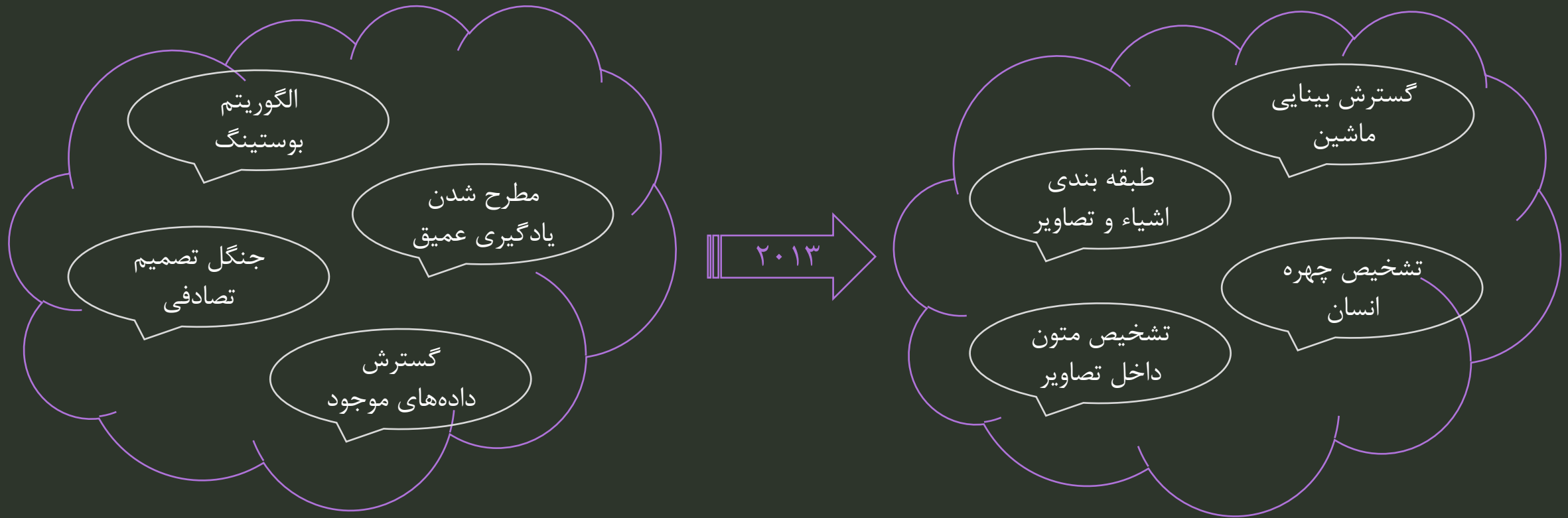


مقدمه

- شبکه‌های عصبی عمیق در مقایسه با عملکرد انسان
- الگوریتم‌های شبکه عصبی عمیق در مقابل حمله‌های خصمانه

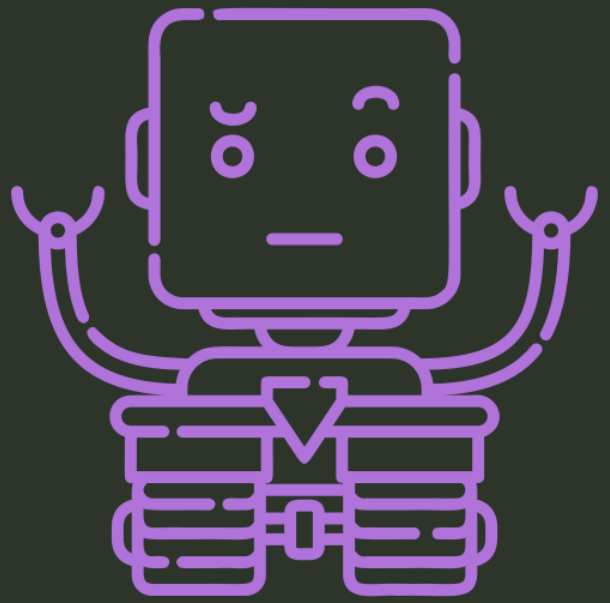
مقدمه

- شبکه‌های عصبی عمیق در مقایسه با عملکرد انسان



مقدمه

- الگوریتم‌های شبکه عصبی عمیق در مقابل حمله‌های خصمانه
  - آیا ماشین نیز می‌تواند اشتباه کند؟
  - این اشتباه‌ها ناشی از چیست؟
  - آیا این اشتباه‌ها توسط انسان قابل درک هستند؟
  - با بروز این مسئله، آیا الگوریتم‌ها نیاز به تغییر دارند؟



اصطلاح‌های مهم

- نمونه اصلی
- اختلال خصمانه
- نمونه خصمانه
- متخاصم
- اختلال نامحسوس
- نرخ فریب
- انتقال پذیری
- حمله هدفمند

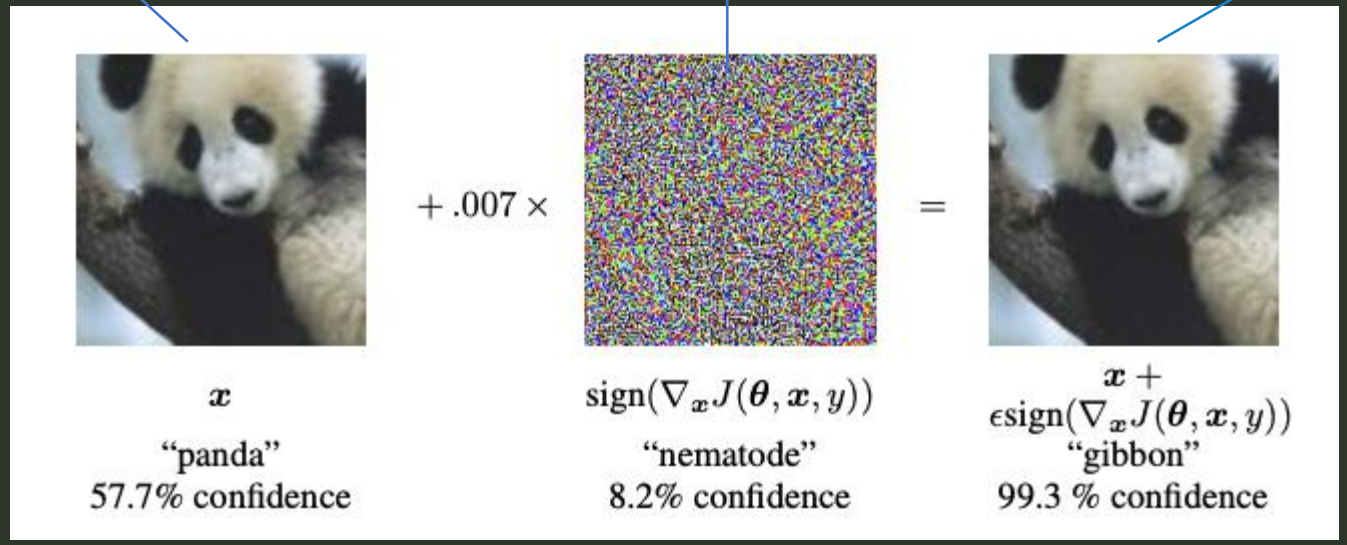
اصطلاح‌های مهم

نمونه/تصویر اصلی

اختلال خصمانه

نمونه/تصویر خصمانه

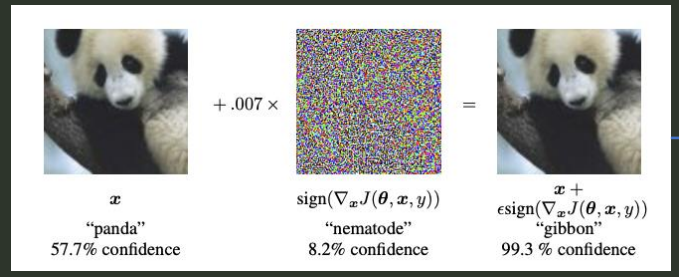
متخاصم

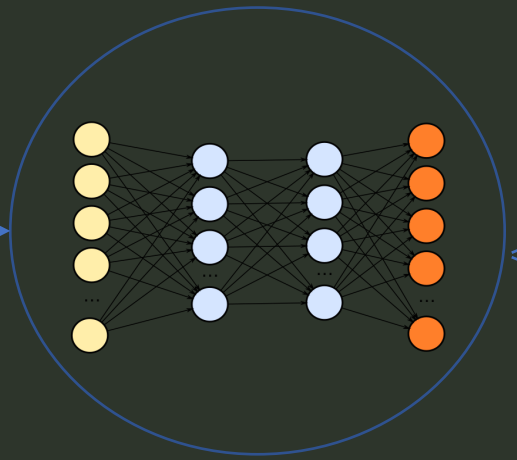


اصطلاح‌های مهم

اختلال نامحسوس

نرخ فریب



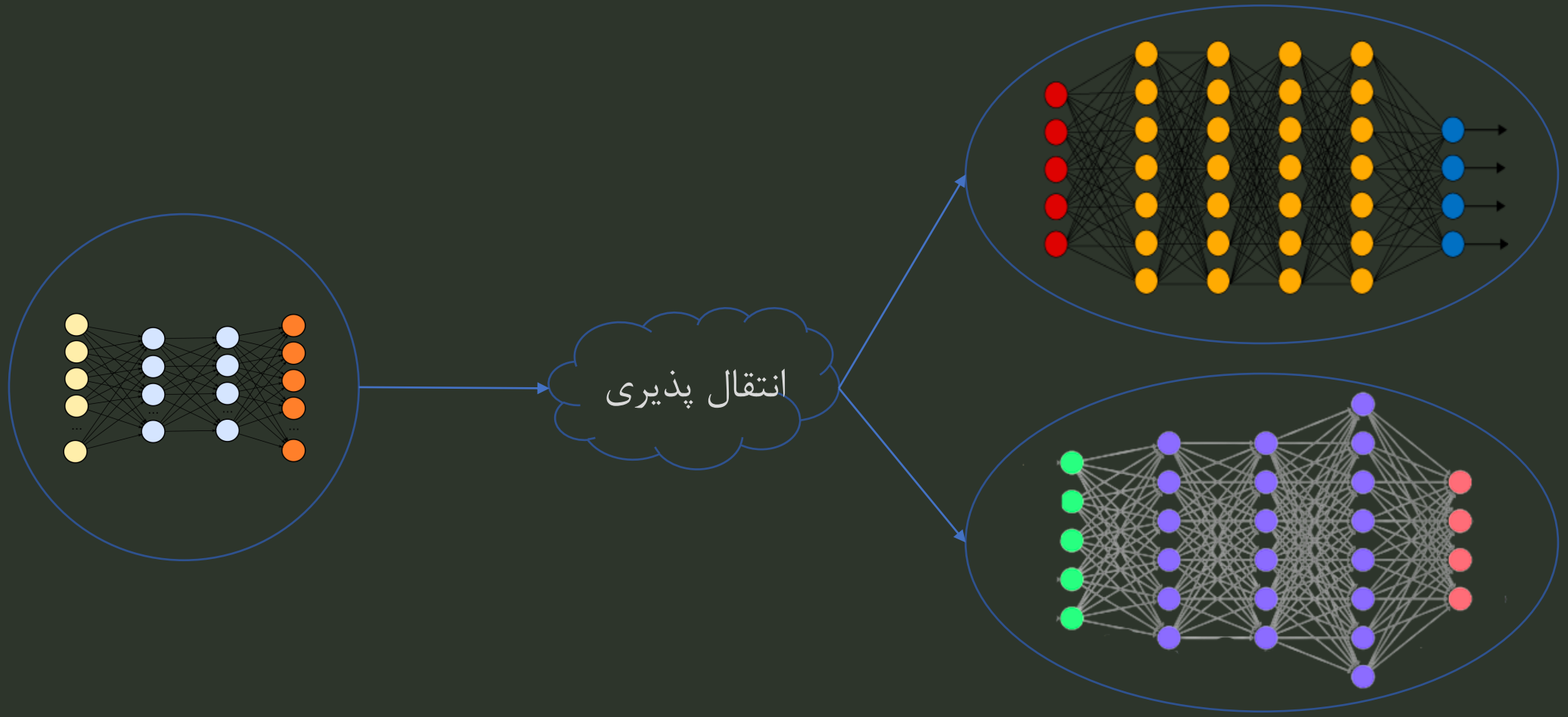


؟

پاندا

حمله هدفمند

اصطلاح‌های مهم

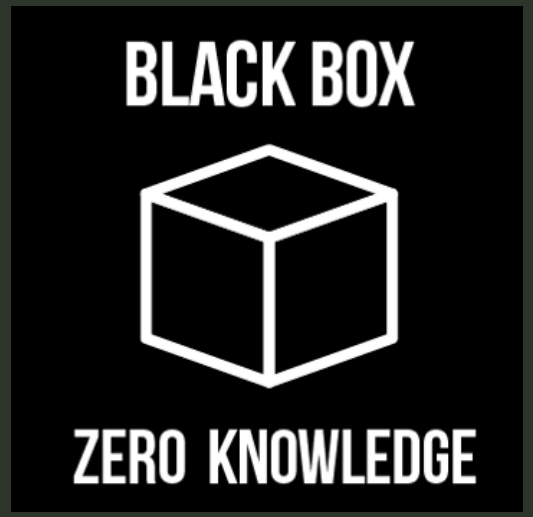


روش‌های تهدید

- حمله‌های جعبه سیاه
- حمله‌های جعبه خاکستری
- حمله‌های جعبه سفید

حمله‌های جعبه سیاه

- اطلاعات حمله کننده:
  - الگوریتم‌های یادگیری عمیق
  - خروجی شبکه برای ورودی‌ها

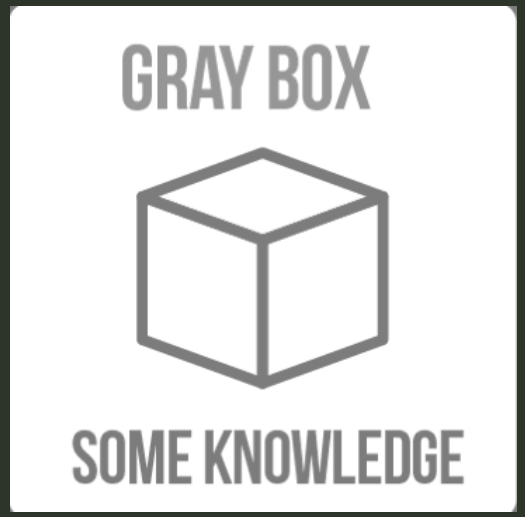


- محدودیت‌های حمله کننده:
  - ساختار معماری شبکه هدف
  - مقادیر متغیرهای شبکه

حمله‌های جعبه خاکستری

- محدودیت‌های حمله کننده:

  - مقادیر متغیرهای شبکه

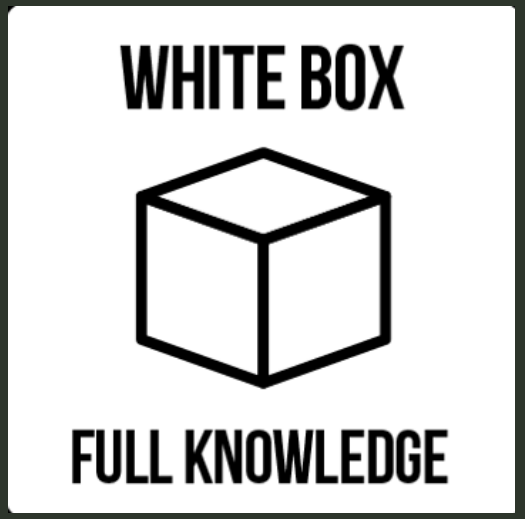


- اطلاعات حمله کننده:

  - الگوریتم‌های یادگیری عمیق
  - خروجی شبکه برای ورودی‌ها
  - ساختار معماری شبکه هدف

حمله‌های جعبه سفید

- محدودیت‌های حمله کننده:



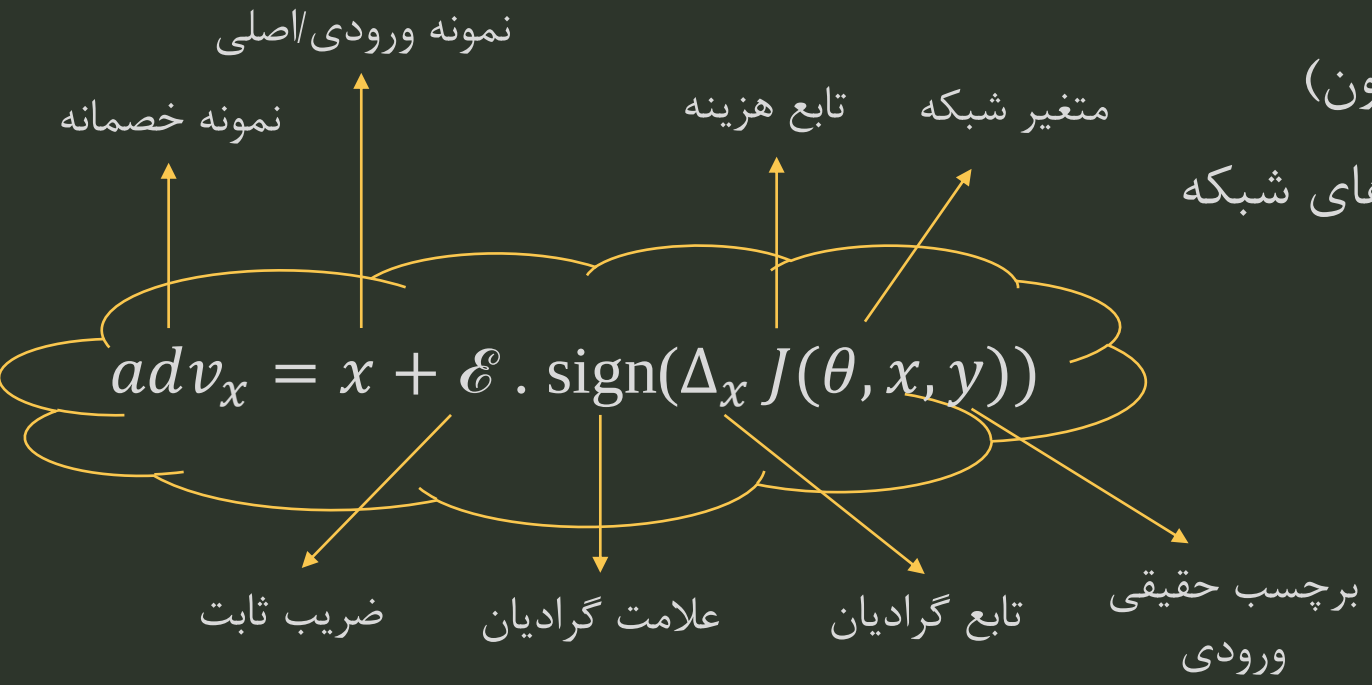
- اطلاعات حمله کننده:
  - الگوریتم‌های یادگیری عمیق
  - خروجی شبکه برای ورودی‌ها
  - ساختار معماری شبکه هدف
  - مقادیر متغیرهای شبکه

الگوریتم‌های حمله

- روش نشانه شیب سریع
- حمله تک-خال

روش نشانه شیب سریع

- حمله جعبه سفید
- اختلال نامحسوس و حداکثری
- نویز خطی روی ورودی اصلی (با ضریب ثابت اپسیلون)
- گرادیان با توجه به تصویر اصلی ورودی و نه متغیرهای شبکه
  - عدم نیاز به تغییر متغیرهای شبکه
  - توقف آموزش مدل



روش نشانه شیب سریع

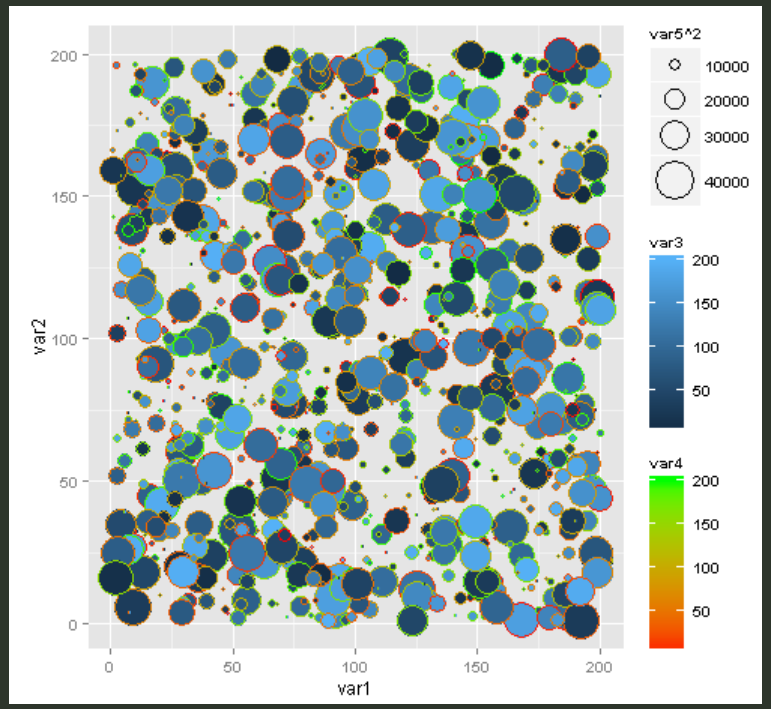
- مثال:

– دقت مدل در داده های آموزشی = ٪ ۹۹.۲۵

– دقت مدل در داده های سنجش = ٪ ۹۸.۷۷

– عملکرد مدل در برابر حمله ؟؟

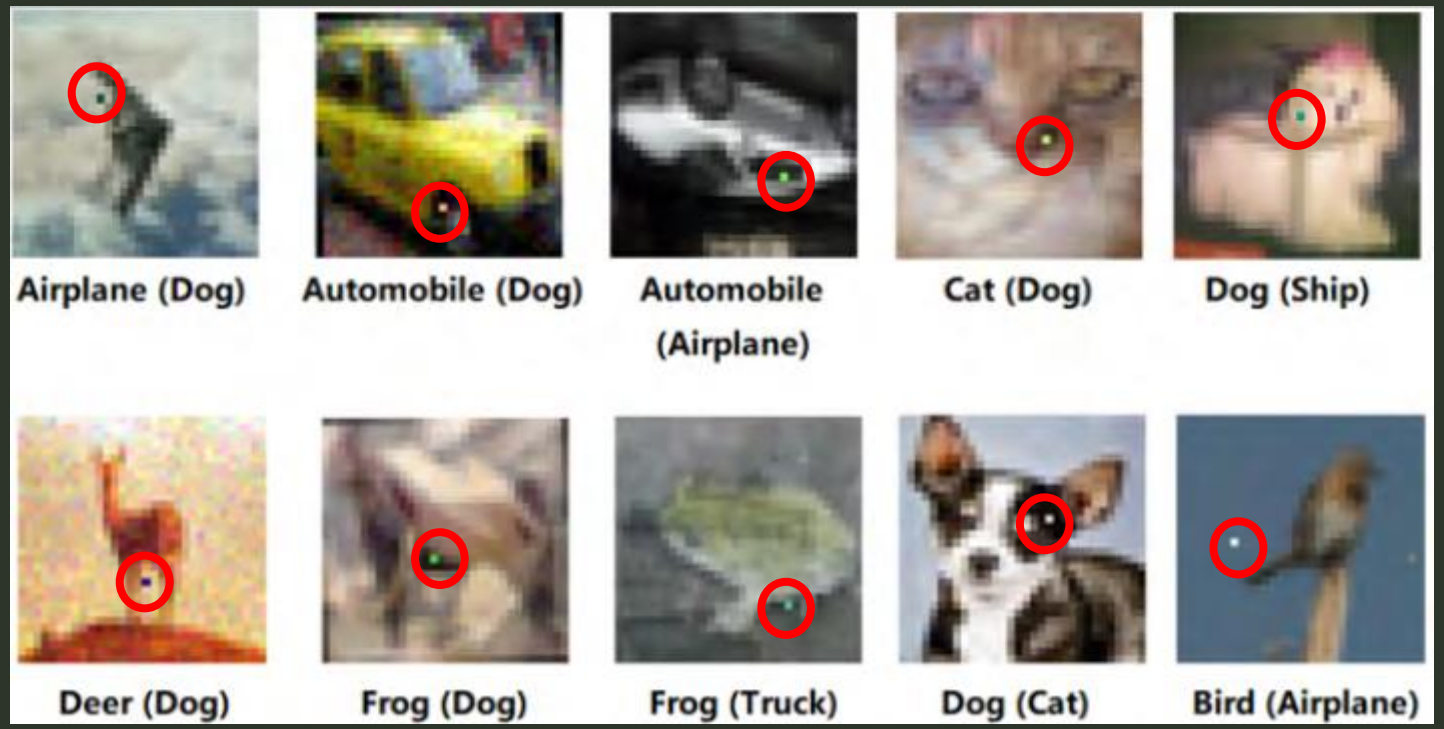
حمله تک-خال



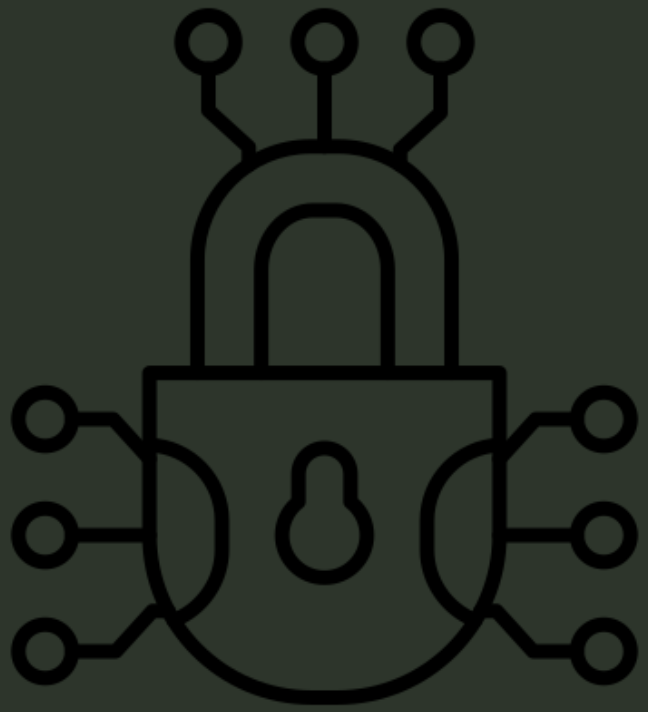
- حمله جعبه سیاه
- استفاده از فضای برداری برای نمونه‌ها/تصاویر
- تغییر تصادفی یک پیکسل (یک یا چند خانه از بردار)
- رقابت فرزند و پدر در تغییر مدل به نفع خود
- حمله غیر هدفمند

حمله تک-خال

- مثال:



تکنیک‌های دفاع



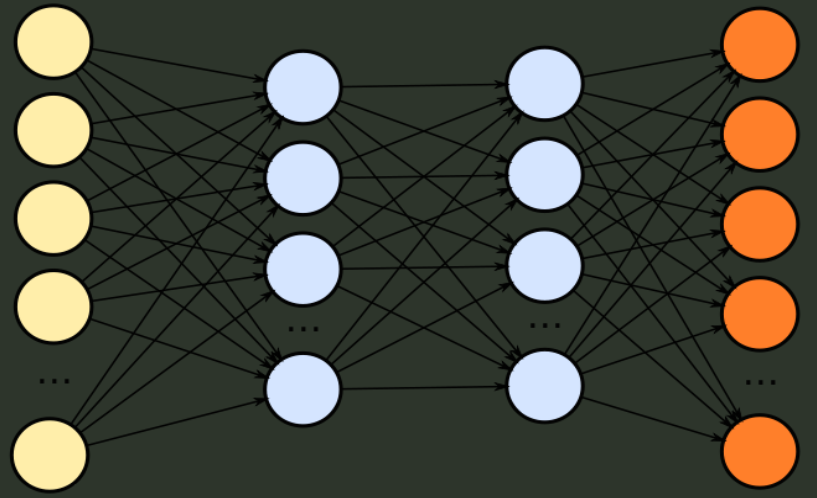
- دسته‌بندی تکنیک‌های دفاع
- آموزش خصمانه

دسته‌بندی تکنیک‌های دفاع

- دفاع تضمین شده
  - کران بالا برای دقت حمله
  - دو عیب بزرگ
    - ۱- مقیاس پذیری ✘
    - ۲- پشتیبانی گسترده ✘
- دفاع ابتکاری
  - مناسب حمله‌های خاص
  - عدم تضمین دقت
  - آموزش خصمانه

آموزش خصمانه

- آموزش شبکه به کمک داده‌های اصلی و داده‌های خصمانه



آموزش خصمانه

- بازی کمینه-بیشینه
  - کمینه کردن مقدار هزینه
  - بیشینه کردن فاصله ورودی و خروجی

$$\min_{\theta} \max_{\mathrm{D}(\mathrm{x},\mathrm{x}')<\mu} J(\theta, x', y)$$

$J(\theta, x', y)$: تابع هزینه خصمانه

θ : وزن شبکه

$\mathrm{D}(\mathrm{x},\mathrm{x}')$: فاصله بین نمونه ورودی و خروجی

x' : نمونه خصمانه

μ : آستانه فاصله بین نمونه ورودی و خروجی

y : خروجی حقیقی

نتیجه گیری

- نتیجه گیری
- پیشنهادات

نتیجه گیری

- جمع‌بندی مطالب ارائه شده
  - مقایسه عملکرد انسان و ماشین
  - اصطلاح‌های مهم در حمله‌های خصمانه
  - انواع روش‌های تهدید
  - تکنیک‌های حمله
  - تکنیک‌های دفاع
- نتیجه ارائه
  - تمرکز بر گسترش الگوریتم‌های مقاوم در برابر حمله

پیشنهادات

- الگوریتم‌های حمله
  - ایجاد شبکه‌ای به منظور آموزش اختلال‌ها (نویز)
  - هدفمند کردن الگوریتم‌های حمله
  - تمرکز بر الگوریتم‌های انتقال‌پذیر
- الگوریتم‌های دفاع
  - گسترش تکنیک‌های دفاع چند جانبه
  - ارزیابی و گسترش الگوریتم‌ها متناسب با محیط فیزیکی
  - تفکر بر تکنیک‌های مقاوم در برابر حمله‌های جعبه سفید

مراجع

[1] Akhtar, Naveed, and Ajmal Mian. "Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey." Ieee Access 6 (2018): 14410-30.

[2] Papernot, Nicolas, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. "Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks." Paper presented at the 2016 IEEE symposium on security and privacy (SP), 2016.

[3] Ren, Kui, Tianhang Zheng, Zhan Qin, and Xue Liu. "Adversarial Attacks and Defenses in Deep Learning." Engineering 6, no. 3 (2020): 346-60.

[4] Xu, Han, Yao Ma, Hao-Chen Liu, Debayan Deb, Hui Liu, Ji-Liang Tang, and Anil K Jain. "Adversarial Attacks and Defenses in Images, Graphs and Text: A Review." International Journal of Automation and Computing 17, no. 2 (2020): 151-78.

[5] Narodytska, Nina, and Shiva Prasad Kasiviswanathan. "Simple Black-Box Adversarial Attacks on Deep Neural Networks." Paper presented at the CVPR Workshops, 2017.

مراجع

[6] Liao, Fangzhou, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. "Defense against Adversarial Attacks Using High-Level Representation Guided Denoiser." Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018.

[7] Madry, Aleksander, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. "Towards Deep Learning Models Resistant to Adversarial Attacks." arXiv preprint arXiv:1706.06083 (2017).

[8] Dong, Yinpeng, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. "Boosting Adversarial Attacks with Momentum." Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018.

[9] Zhang, Yuheng, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. "The Secret Revealer: Generative Model-Inversion Attacks against Deep Neural Networks." Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[10] Carlini, Nicholas, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. "On Evaluating Adversarial Robustness." arXiv pre-print server (2019-02-20 2019). https://doi.org/None
arxiv:1902.06705.

لطفا سوال‌های خود را مطرح بفرمایید

با تشکر از توجه شما

zandiyeh1379@gmail.com