به نام خدا

تکلیف سوم روش پژوهش و ارائه

مطالعه منابع و یادداشت برداری

محمد جواد زندیه ۹۸۳۱۰۳۲

۱۹ اسفند ۱۴۰۰

دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر

بخش اول

مقاله انتخاب شده برای مطالعه

شماره مقاله: ۳

عنوان مقاله: ***"Adversarial Attacks and Defenses in Deep Learning"***

ساختار اولیه پژوهش

۱- مقدمه

۲- مفاهیم پایه

۲-۱ تعاریف

۲-۱-۱ اصطلاحات مهم

۲-۱-۲ نماد گذاری ها و ماتریس ها

۲-۲ مدل های تهدید

۲-۲-۱ جعبه سیاه (black box)

۲-۲-۲ جعبه خاکستری (semi-white(gray) box)

۲-۲-۳ جعبه سفید (white box)

۳- حملات خصمانه

۳-۱ معرفی تعدادی از حوزه های مورد حمله

۳-۱-۱ بینایی کامپیوتر

۳-۱-۲ پردازش تصویر

۳-۱-۳ ساختارهای گرافی

۳-۱-۴ پردازش متن

۳-۱-۵ تشخیص گفتار

۳-۲ الگوریتم های حمله

۴- دفاع در برابر حملات خصمانه

۴-۱ دسته بندی تکنیک های دفاع

۴-۱-۱ ابتکاری (heuristic)

۴-۱-۲ تضمین شده (certificated)

۴-۲ الگوریتم های دفاع

۴-۳ ارزیابی الگوریتم های دفاع

۵- بحث ها

۵-۱ چالش های اساسی حل نشده

۵-۲ چشم انداز های تحقیقاتی

۶- نتیجه گیری

مراجع

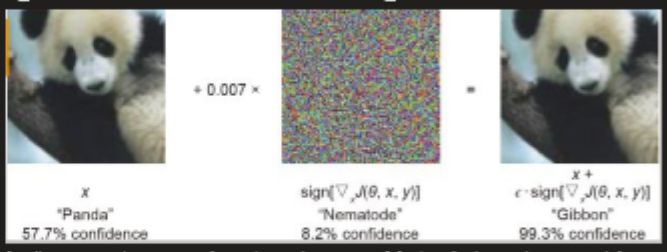
بخش دوم

فیش ها

برای تهیه فیش ها از OneNote استفاده شده است که در فایل 03-fiches تمامی فیش ها قرار داده شده است و برای نمونه هم یکی از فیش ها را در زیر مشاهده می کنید.

شماره فیش	بخش مربوطه	شماره منبع	صفحه منبع	نوع یادداشت
۱	۱- مقدمه	۳	۳۴۶ (۱ / ۱۵)	نقل غیر مستقیم

- با گسترش AI و DL آسیب پذیری امنیتی این حوزه ها نیز در قبال حملات خصمانه مطرح شده است.
- گسترش DL به علت قدرت محاسباتی بالا در این حوزه می باشد که توانسته است مسائل مختلفی در ML را برطرف کند.
- این حملات طوری می توانند عمل کنند که از دید یک ناظر انسانی قابل درک نباشند اما به راحتی اثرات مخرب خود را بر DL می گذارند.
- اثر مخرب این حملات به حدی می تواند باشد که ماشین ما با درصد بسیار بالایی بر غلط خود پافشاری کند.(تصویر)



در شکل سمت راست با دقت ۹۹.۳ درصد مطمئن است که میمون است در حالی که قبل پیش از نویز و در تصویر سمت چپ با درصد ۵۷.۷ میگفت که پاندا است. در واقع این تصویر میزان اثر گذاری بسیار زیاد حملات را نشان میدهد.

- کاربردی بودن بررسی این موضوع در حملات خصمانه ای که در دنیای فیزیکی رخ میدهد قابل مشاهده است.
- بررسی حملات و دفاع در هر دو حوزه یادگیری ماشین و امنیت به موضوعی داغ و به روز برای تحقیق در حال حاضر تبدیل شده است.

نکات:

- بررسی حوزه های مورد حمله در دنیای فیزیکی در قسمت ۳-۱ انجام می شود.
- ML - machine learning
- DL - deep learning

مقاله مطالعه شده

مقاله مطالعه شده با نام 03- Adversarial Attacks and Defenses in Deep Learning-commented در کنار این فایل قرار داده شده است و برای نمونه هم بخشی از قسمت مطالعه شده در اینجا قرار داده می شود.

In Section 6, we provide some observations and insights on several related open problems. In Section 7, we conclude this survey.

مشابه قسمت بالا باید برای مقدمه تحقیق انجام شود.

2. Preliminaries

۲-۱-۲

2.1. Definitions and notations

In this section, we first clarify the definitions and notations used in this paper. Specifically, a dataset is defined as $\{x_i, y_i\}_{i=1}^{N}$, where x_i is a data sample with label y_i, and N is the size of the dataset. A neural network is denoted as $f(\cdot)$ with input x and prediction $f(x)$. The corresponding optimization loss (also called adversarial loss) is denoted by $J(\theta, x, y)$, where θ represents the model weights. For a classification task, the cross-entropy between $f(x)$ and the label (one-hot) y is always applied as the optimization loss, which is denoted by $J(f(x), y)$. A data sample x' is considered to be an adversarial sample of x when x' is close to x under a specific distance metric while $f(x') \neq y$. Formally, an adversarial sample of x is defined as follows:

$$x' : D(x, x') < \eta,\ f(x') \neq y \tag{1}$$

where $D(\cdot,\cdot)$ is the distance metric and η is a predefined distance constraint, which is also known as the allowed perturbation. Empirically, a small η is adopted to guarantee the similarity between x and x' such that x' is indistinguishable from x.

2.2. Distance metrics

By definition, an adversarial sample x' should be close to a benign sample x under a specific distance metric. The most commonly used distance metric is the L_p distance metric [8]. The L_p distance between x and x' is denoted as $||x - x'||_p$, where $||\cdot||_p$ is defined as follows:

$$||\boldsymbol{v}||_p = \left(|\boldsymbol{v}_1|^p + |\boldsymbol{v}_2|^p + \cdots + |\boldsymbol{v}_d|^p\right)^{1/p} \tag{2}$$

۲-۱-۲

information, a gray-box adversary always shows better attack performance compared with a black-box adversary. The strongest adversary—that is, the white-box adversary—has full access to the target model including all the parameters, which means that the adversary can adapt the attacks and directly craft adversarial samples on the target model. Currently, many defense methods that have been demonstrated to be effective against black-box/gray-box attacks are vulnerable to an adaptive white-box attack. For example, seven out of nine heuristic defenses in the 2018 International Conference on Learning Representations (ICLR2018) were compromised by the adaptive white-box attacks proposed in Ref. [17].

3. Adversarial attacks

این بخش به طور وسیعی در ۲-۳ مورد استفاده قرار میگیرد

In this section, we introduce a few representative adversarial attack algorithms and methods. These methods target to attack image classification DL models, but can also be applied to other DL models. We detail the specific adversarial attacks on the other DL models in Section 4.

3.1. L-BFGS algorithm

The vulnerability of deep neural networks (DNNs) to adversarial samples is first reported in Ref. [4]; that is, hardly perceptible adversarial perturbations are introduced to an image to mislead the DNN classification result. A method called L-BFGS is proposed to find the adversarial perturbations with the minimum L_p norm, which is formulated as follows:

$$\min_{x} ||x - x'||_p \text{ subject to } f(x') \neq y' \tag{3}$$

where $||x - x'||_p$ is the L_p norm of the adversarial perturbations and y' is the adversarial target label ($y' \neq y$). However, this optimization problem is intractable. The authors propose minimizing a hybrid