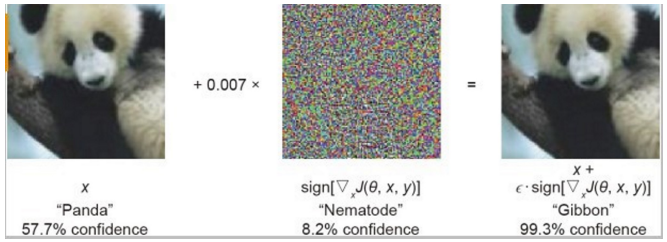
فیش ۱

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

فیش ۲

شماره فیش	بخش مربوطه	شماره منبع	صفحه منبع	نوع یادداشت
۲	۲-۲-۱ جعبه سیاه (black box)	۳	(۱۵ / ۱) ۳۴۶ و (۱۵ / ۲) ۳۴۷	نقل غیر مستقیم

- حمله کننده ها ساختار معماری شبکه هدف و پارامتر ها را نمیداند و فقط می تواند با الگوریتم DL در ارتباط باشد.
- تنها راه نفوذ در این مدل، کوئری و درخواست زدن به منظور پیش بینی الگوریتم DL در قبال ورودی های خاص باشد.
- حمله کننده ها همواره نمونه های خصمانه را بر روی یک classifier جایگزین -که توسط جفت های داده و پیش بینی و دیگر نمونه های خصمانه پنهان از ناظر انسانی آموزش دیده شده اند- می سازند.
- حملات جعبه سیاه همواره می توانند مدل هایی که به صورت طبیعی در برابر حملات تجهیز و آموزش دیده نشده اند را به خطر بیندازند.

نکات:

- تفاوت مدل های تهدید جعبه سفید و خاکستری و سیاه در سطح دانش حمله کنندگان نسبت به مدل هدف (معماری و پارامتر های مدل) می باشد .
- حمله کننده = adversary
- اطلاعات کامل در مورد این بخش در مقاله شماره ۵ قرار دارد که به طور مفصل بحث خواهد شد.

فیش ۳

شماره فیش	بخش مربوطه	شماره منبع	صفحه منبع	نوع یادداشت
۳	۲-۲-۲ جعبه خاکستری (semi-white(gray) box)	۳	۳۴۶ (۱ / ۱۵) و ۳۴۷ (۲ / ۱۵)	نقل غیر مستقیم

- اطلاعات حمله کننده محدود به ساختار مدل هدف است و در مورد پارامتر ها و وزن ها اطلاعاتی ندارد.
- همانند مدل تهدید جعبه سیاه، تنها راه نفوذ در این مدل، کوئری و درخواست زدن به منظور پیش بینی الگوریتم DL در قبال ورودی های خاص می باشد.
- برخلاف جعبه سیاه، حمله کننده نمونه های خصمانه خود را روی یک مدل جایگزین با همان معماری مدل اصلی بنا می کند (زیرا گفتیم که معماری مدل هدف را میداند)
- اینگونه حملات نسبت به مدل جعبه سیاه از اطلاعاتی بیشتری نسبت به مدل هدف برخوردار هستند پس انتظار می رود که کارایی بیشتری داشته باشند.

نکات:

- تفاوت مدل های تهدید جعبه سفید و خاکستری و سیاه در سطح دانش حمله کنندگان نسبت به مدل هدف (معماری و پارامتر های مدل) می باشد .
- حمله کننده = adversary
- در مقالات آتی در مورد مدل های تهدید بیشتر اطلاعات کسب خواهیم کرد.

فیش ۴

شماره فیش	بخش مربوطه	شماره منبع	صفحه منبع	نوع یادداشت
۴	۲-۲-۳ جعبه سفید (white box)	۳	۳۴۶ (۱ / ۱۵) و ۳۴۷ (۲ / ۱۵)	نقل غیر مستقیم

- حمله کننده ها اطلاعات کامل از مدل هدف یعنی معماری و پارامتر های مدل دارند.
- حمله کننده ها می توانند به طور مستقیم نمونه های متخاصم را در مدل هدف به هر وسیله ای ایجاد کنند، زیرا اطلاعات کامل از آن دارند.
- انواع الگوریتم های حمله طراحی شده بر اساس مدل جعبه سفید:

ple generation have been proposed, such as limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) algorithm [4], the fast gradient sign method (FGSM) [5], the basic iterative method (BIM)/projected gradient descent (PGD) [6], distributionally adversarial attack [7], Carlini and Wagner (C&W) attacks [8], Jacobian-based saliency map attack (JSMA) [9], and DeepFool [10]. These

- اکثر الگوریتم های دفاعی موجود در حال حاضر در قبال این مدل تهدید آسیب پذیرند و به نوعی فلج می باشند.

نکات:

- تفاوت مدل های تهدید جعبه سفید و خاکستری و سیاه در سطح دانش حمله کنندگان نسبت به مدل هدف (معماری و پارامتر های مدل) می باشد
- حمله کننده = adversary
- در مقالات آتی در مورد مدل های تهدید بیشتر اطلاعات کسب خواهیم کرد.

فیش ۵

شماره فیش	بخش مربوطه	شماره منبع	صفحه منبع	نوع یادداشت
۵	۴-۱ دسته بندی تکنیک های دفاع	۳	۳۴۶ (۱ / ۱۵) و ۳۴۷ (۲ / ۱۵)	نقل غیر مستقیم

- دو تکنیک کلی برای دفاع در برابر حملات خصمانه وجود دارد: ۱. ابتکاری (heuristic) ۲. تضمین شده (certificated)
- دسته اول: تکنیک های ابتکاری
  ۱. این تکنیک ها موقعی خوب عمل می کنند که حملات خاصی مورد نظر باشد و تضمینی هم برای دقت عملکرد دفاعی به صورت تئوری مورد نیاز نباشد.
  ۲. در حال حاضر موفق ترین دفاع ابتکاری، تکنیک آموزش خصمانه یا adversarial training می باشد.
  ۳. در adversarial training نمونه های متخاصم را با نمونه های دیگر در مرحله آموزش ترکیب می کنیم.
  ۴. از دسته adversarial training ها الگوریتم PGD بهترین دقت در حال حاضر را دارا می باشد.
  ۵. دسته دیگری از تکنیک های ابتکاری به حذف نویز و تبدیل ورودی ها / ویژگی ها برای کاهش اثر حملات روی داده ها / ویژگی ها وابسته اند.
- دسته دوم: تکنیک های تضمین شده
  ۱. این دسته از تکنیک ها همواره یک کلاس خوش-تعریف مجزا برای حملات خصمانه در نظر گرفته اند.
  ۲. با فرموله کردن یک polytope متخاصم، کران بالای میزان حملات را بدست می آورند.
  ۳. در واقع این روش تضمین می کند که هیچ حمله ای نمیتواند با دقتی فراتر از کران بالای تخمین زده شده رخ دهد.
- مقایسه دو تکنیک: کارایی تکنیک اول و روش adversarial training از تکنیک های certificated بسیار بیشتر است.

نکات:

- ویژگی = feature
- خوش-تعریف = well-defined
- تکنیک دوم (تکنیک های تضمین شده) خیلی سخت توضیح داده شده است و باید مورد بررسی بیشتر قرار بگیرد.

فیش ۶

شماره فیش	بخش مربوطه	شماره منبع	صفحه منبع	نوع یادداشت
۶	۲-۱-۲ نماد گذاری ها و ماتریس ها	۳	۳۴۷ (۲ / ۱۵)	نقل مستقیم

مجموعه داده dataset	شبکه عصبی Neural network	خطای بهینه سازی Optimization loss	cross-entropy بین f(x) و y	نمونه خصمانه Adversarial sample
$\{x_i, y_i\}_{i=1}^{N}$	$f(x)$	$J(\theta, x, y)$	$J(f(x); y)$	x'

نمونه داده Data sample	لیبل نمونه داده Data sample label	سایز مجموعه داده Dataset size	فرمول نمونه خصمانه Adversarial sample formula
x_i	y_i	N	$x' : D(x, x') < \eta, f(x') \neq y$

ماتریس فاصله Distance metric	محدوده فاصله از پیش تعیین شده	ماتریس فاصله بین x و x'	تعریف ماتریس فاصله
$D(\cdot, \cdot)$	η	L_p = $\|\|x - x'\|\|_p$	$\|\|\boldsymbol{v}\|\|_p = (\|\boldsymbol{v}_1\|^p + \|\boldsymbol{v}_2\|^p + \cdots + \|\boldsymbol{v}_d\|^p)^{1/p}$

عدد حقیقی	بعد ماتریس فاصله v
p	d

نکات: برای کسب اطلاعات بیشتر به صفحه۲ مقاله مراجعه شود.

فیش ۷

شماره فیش	بخش مربوطه	شماره منبع	صفحه منبع	نوع یادداشت
۷	۳-۲ الگوریتم های حمله	۳	۳۴۷ (۲ / ۱۵) تا ۳۵۱ (۶ / ۱۵)	نقل مستقیم

- در این فیش به صورت خلاصه اسامی الگوریتم های حمله ای که در مقاله ذکر شده اند را بیان می کنیم و توضیحات بیشتر در هر مورد را با توجه به علامت هایی که در مقاله گذاشته شده است می توان بدست آورد و از بیان مجدد آنها صرف نظر می کنیم (خود مقاله به صورت خلاصه هر یک از الگوریتم ها را بیان کرده و لازم به ذکر مجدد نیست)
- انواع الگوریتم های حمله

۱. L-BFGS algorithm
۲. Fast gradient sign method
۳. BIM and PGD
۴. Momentum iterative attack
۵. Distributionally adversarial attack
۶. Carlini and Wagner attack
۷. Jacobian-based saliency map approach
۸. DeepFool
۹. Elastic-net attack to DNNs
۱۰. Universal adversarial attack
۱۱. Adversarial patch
۱۲. GAN-based attacks
۱۳. Practical attacks
۱۴. Obfuscated-gradient circumvention attacks

نکات:

- نکات هر یک از الگوریتم ها به صورت علامت گذاری شده در مقاله وجود دارد.

فیش ۸

شماره فیش	بخش مربوطه	شماره منبع	صفحه منبع	نوع یادداشت
۸	۲-۴ الگوریتم های دفاع	۳	(۱۵ / ۷) ۳۵۲ تا (۱۵ / ۱۲) ۳۵۷	نقل غیر مستقیم

• در این فیش انواع مدل های دفاع را به صورت موردی بیان می کنیم و جزئیات آنها در داخل مقاله علامت گذاری شده است:

۱. Adversarial training
۱.۱ FGSM adversarial training
۱.۲ PGD adversarial training
۱.۳ Ensemble adversarial training
۱.۴ Adversarial logit pairing
۱.۵ Generative adversarial training
۲. Randomization
۲.۱ Random input transformation
۲.۲ Random noising
۲.۳ Random feature pruning
۳. Denoising
۴. Provable defenses
۵. Weight-sparse DNNs
۶. KNN-based defenses
۷. Bayesian model-based defenses
۸. Consistency-based defenses

نکات:

فیش ۹

شماره فیش	بخش مربوطه	شماره منبع	صفحه منبع	نوع یادداشت
۹	۵ بحث ها	۳	۳۵۷ (۱۳ / ۱۵) و ۳۵۸ (۱۳ / ۱۵)	نقل غیر مستقیم

- روند تحقیق در این حوزه شامل دو بخش است:
  ۱. طراحی حملات کارآمد تر به منظور ارزیابی و تقویت الگوریتم های دفاعی
  ۲. تحقیق در حوزه حملات در دنیای فیزیکی
- حملات خصمانه ای که در دنیای دیجیتال (digital space) طراحی شده اند در دنیای فیزیکی کارآمد نیستند زیرا شرایط محیطی در آنها در نظر گرفته نشده است.
- تمرکز حال حاضر روی الگوریتم های دفاعیِ "certificated" است زیرا اکثر الگوریتم های دفاعیِ "heuristic" در برابر تهدید های مدل جعبه سفید بسیار آسیب پذیری هستند.
- مشکل الگوریتم های دفاعیِ certificated مقیاس ناپذیری آنها است.
- در حال حاضر توسعه الگوریتم های دفاعی بیش از الگوریتم های حمله مورد چالش است زیرا جامعه هدف آنها بیشتر از الگوریتم های حمله است.
- تعدادی از چالش های حل نشده در این حوزه تحقیقاتی تا ژوئن سال ۲۰۲۰ که تاریخ چاپ این مقاله می باشد:
  ۱. علیت پشت نمونه های متخاصم هنوز مشخص نیست: تحقیقات اخیر نشان می دهد که آسیب پذیری در برابر نمونه های متخاصم احتمالا به علت بُعد بالای مجموعه داده ها (یعنی تعداد فیچر های زیاد) و عدم آموزش مدل با استفاده از داده های آموزشی کافی می باشد.
  ۲. وجود یک کران بالا و مرز برای تعیین میزان قدرت مدل در برابر حملات متخاصم؟
  ۳. یک الگوریتم کارا و با دقت بالا در برابر تهدید های نوع جعبه سفید وجود دارید؟ از لحاظ کارایی الگوریتم هایی بیان شده است، از لحاظ بهره وری و دقت نیز الگوریتم هایی بیان شده است اما الگوریتمی که هر دو را با هم در برابر این نوع تهدیدات داشته باشند هنوز دور از ذهن به نظر می رسد.

نکات:

- مطالب این فیش سرنخ های تحقیقاتی آینده در این حوزه میباشد.
- بررسی شود که آیا مطالب حل نشده تا تاریخ انتشار این مقاله آیا اکنون حل شده است یا نه و ....

فیش ۱۰

شماره فیش	بخش مربوطه	شماره منبع	صفحه منبع	نوع یادداشت
۱۰	۶ نتیجه گیری	۳	۳۵۸ (۱۳ / ۱۵)	نقل غیر مستقیم

• مباحثی که در این مقاله مطرح شده است به صورت خلاصه:

۱. انواع الگوریتم های حمله و دفاع

۲. بررسی کارایی الگوریتم های دفاع

۳. بررسی الگوریتم های جدید حمله و دفاع

۴. بررسی مسائل پایه ای و تعاریف

۵. اثبات اینکه در حال حاضر هیچ الگوریتم دفاعی وجود ندارد که هم کارا باشد و هم هزینه کمی داشته باشد

۶. تکنیک adversarial training کاراست اما هزینه زیادی دارد

۷. تکنیک های دفاعی heuristic در مقابل تهدید های مدل جعبه سفید آسیب پذیر است

۸. در مورد مسائل و چالش های حل نشده این حوزه بحث شده است

۹. سرنخ های تحقیقاتی در این زمینه بررسی شده است

نکات: