به نام خدا

دانشکده مهندسی کامپیوتر

امنیت و حریم خصوصی در یادگیری ماشین

(۴۰۸۱۶)(نیم‌سال دوم سال تحصیلی ۱۴۰۱-۱۴۰۲)

استاد درس: دکتر امیرمهدی صادق‌زاده

دستیاران آموزشی: مهدی غزنوی، زینب گلگونی، الهه فرشادفر، محمدرضا کاظمی، حمید دشتبانی

تمرین چهارم

مهلت تحویل: ساعت ۲۳:۵۹ شنبه ۳۰ اردیبهشت ۱۴۰۲

نکات و قواعد

۱. سوالات خود را زیر پیام مربوطه در Quera مطرح نمایید.

۲. محل بارگذاری تمرین تا یک هفته پس از مهلت ارسال باز خواهد بود. در طول ترم، در مجموع می‌توانید از ۲۴ روز تاخیر مجاز به صورت ساعتی استفاده کنید و پس از آن به ازای هر روز ۲۰ درصد جریمه بر روی نمره‌ی کسب شده اعمال خواهد شد.

۳. لطفا مطابق تاکید پیشین، حتما آداب‌نامه‌ی انجام تمرین‌های درسی را رعایت نمایید. در صورت تخطی از آیین‌نامه، در بهترین حالت مجبور به حذف درس خواهید شد.

۴. در صورتی که پاسخ‌های سوالات نظری را به صورت دست‌نویس آماده کرده‌اید، لطفا تصاویر واضحی از پاسخ‌های خود ارسال کنید. در صورت ناخوانا بودن پاسخ ارسالی، نمره‌ای به پاسخ ارسال شده تعلق نمی‌گیرد.

۵. همه‌ی فایل‌های مربوط به پاسخ خود را در یک فایل فشرده و با نام SPML_HW۴_StdNum_FirstName_LastName ذخیره کرده و ارسال نمائید.

سوال ۱ آشفتگی‌های خصمانه‌ی فراگیر (۴۰ نمره)

(الف) (۴ نمره) آشفتگی‌های تولید شده توسط الگوریتم UAP به چه معنا فراگیر هستند؟ آیا آشفتگی فراگیر به دست آمده برای یک مدل قابلیت تعمیم به سایر مدل‌ها را دارد؟ اگر پاسخ بله است، شیوه‌ی کارکرد الگوریتم را توضیح دهید و اگر پاسخ خیر است، در یک مثال کوچک [۱] این تعمیم نیافتن را توضیح دهید.

(ب) (۳ نمره) یافتن آشفتگی خصمانه‌ی فراگیر چه اهمیت و کاربردی دارد؟

(ج) (۳ نمره) با داشتن دادگان D و تابع g که میزان موفقیت حمله را اندازه می‌گیرد (هر چه $g(x)$ بالاتر باشد، موفقیت نمونه‌ی خصمانه‌ی x بیشتر است)، یافتن آشتفگی خصمانه‌ی فراگیر را به صورت یک مسئله‌ی بهینه‌سازی مقید به کرانی بر روی نرم [۲] آشفتگی بنویسید.

(د) (۳۰ نمره) در این قسمت با کمک دفترچه‌ی UAP.ipynb و فایل‌های کمکی که در اختیار شما قرار گرفته است، الگوریتم UAP را پیاده سازی خواهید کرد. برای این کار بخش‌های لازم را در کد تکمیل کنید و تابع مربوط به الگوریتم را پیاده‌سازی کنید.

برای مشاهده قابلیت تعمیم نمونه‌ی فراگیر، نحوه کار به این صورت خواهد بود که مجموعه دادگان ارزیابی CIFAR10 به دو بخش مساوی تقسیم می‌شوند. یک بخش برای یادگیری آشفتگی فراگیر به عنوان ورودی به تابع داده می‌شود (دادگان مهاجم). بخش دیگر برای ارزیابی درصد موفقیت نمونه خصمانه فراگیر بعد از اجرای تابع استفاده می‌شود (دادگان ارزیابی نهایی).

پس از پیاده‌سازی تابع خود، آن را برای دادگان مهاجم که شامل نصف دادگان ارزیابی CIFAR10 است، فراخوانی کنید.

نمودار مربوط به درصد موفقیت آشفتگی فراگیر روی دادگان مهاجم در هر دور از الگوریتم را رسم کنید.

[۱] Toy example
[۲] Norm

آشفتگی فراگیر خروجی الگوریتم خود را به عنوان نمونه خصمانه بر روی مجموعه دادگان ارزیابی نهایی CIFAR10 (که مهاجم به آن‌ها دسترسی نداشته است) استفاده کنید و درصد موفقیت آن را گزارش کنید.

سوال ۲ حملات خصمانه‌ی فیزیکی (۲۰ نمره)

در دنیای واقعی نیز نمونه‌های خصمانه‌ای یافت می‌شوند که می‌توانند سبب اختلال کارکرد مدل‌های مبتنی بر یادگیری ماشین شوند.

(الف) (۵ نمره) وصله‌های خصمانه نمونه‌ای از حملات خصمانه‌ی فیزیکی هستند که با آن در کلاس آشنا شدید. در همین رابطه، با مطالعه‌ی چکیده و مقدمه‌ی مقاله‌ی Adversarial Board به صورت کلی هدف و روش استفاده شده در این مقاله را توضیح دهید.

(ب) (۵ نمره) یک چالش مهم برای اعمال حملات خصمانه‌ای که تاکنون آموخته‌ایم در دنیای واقعی آن است که نمونه‌های خصمانه‌ی دیجیتالی را نمی توان به طور دقیق در فضای فیزیکی تحقق بخشید. مثالی از این مورد بزنید. چه راه حل‌هایی برای برطرف کردن اینگونه مشکلات وجود دارد؟

(ج) (۵ نمره) چالش دیگر آن است که ممکن است این نمونه‌ها به خوبی به محیط‌های فیزیکی تعمیم نیابند. مثالی از این مورد بزنید. چه راه حل‌هایی برای برطرف کردن اینگونه مشکلات وجود دارد؟

(د) (۵ نمره) Expectation over transformation روشی برای ساخت نمونه‌ی خصمانه‌ی مقاوم x از روی نمونه داده‌ی x_0 تحت تبدیلات [۳] $\mathcal{T}$ معرفی می‌کند. یافتن این نمونه‌ی خصمانه را به صورت یک مسئله‌ی بهینه‌سازی بنویسید.

سوال ۳ ارزیابی مقاومت روش‌های دفاع (۳۰ نمره)

همان‌طور که در درس مطرح شد، با شناسایی نمونه‌های خصمانه، تلاش برای دفاع دربرابر حملات و مقاوم‌سازی مدل‌ها به یک مسئله جدی تبدیل شده‌است. روش‌های مختلفی نیز در این سال‌ها پیشنهاد شده‌اند. نکته مهم در این میان، ارزیابی جامع و قابل اطمینان موفقیت روش‌های دفاع پیشنهادی است.

(الف) (۱۵ نمره) یک تیم پژوهشی روش جدیدی برای بهبود مقاومت مدل‌ها نسبت به نمونه‌های خصمانه به نام prpd پیشنهاد کرده‌اند. در این روش تابع Relu با تابع دیگری جایگزین شده است و ادعا می‌شود در مقابل حملات مختلف مقاومت پایدار بالاتری دارد. این تیم برای اثبات اثربخشی روش خود عملکرد آن را در مقابل حملات FGSM، PGD-20، CW و همچنین BB که یک حمله جعبه سیاه نسبتا ساده مبتنی بر مدل دیگر است، گزارش کرده‌اند که نتایج آن در جدول زیر نمایش داده شده‌است.

BB	CW	PGD-20	FGSM	Clean	
46.93	47.12	47.25	47.62	81.40	prpd

نظر شما در باره ارزیابی درستی ادعای موفقیت این روش پیشنهادی چیست؟ آیا در همین نتایج و ارزیابی صورت گرفته، نکته یا نکات سوال‌برانگیزی وجود دارد که موفقیت این روش را برای شما زیر سوال ببرد؟ تحلیل خود را بیان کنید و همچنین در صورتی که به نظرتان نیاز به بررسی بیشتر وجود دارد، دو راه برای ادامه ارزیابی این روش دفاع پیشنهاد کنید.

(ب) (۱۵ نمره) یکی از نتایجی که در کارهای اخیر برای ارزیابی مقاومت روش‌های پیشنهادی گزارش می‌شود، دقت خصمانه در مقابل مجموعه‌ای از حملات موسوم به AutoAttack است.

مقاله Reliable Evaluation of Adversarial Robustness with an Ensemble of Diverse Parameter-free Attacks مجموعه‌ای از حملات متنوع را در کنار هم معرفی و پیاده‌سازی کرده است که با توجه به قدرت و اثربخشی آن در کارهای بعدی به عنوان معیار شناخته شده‌ای برای ارزیابی مقاومت مدل‌ها استفاده شده است.

هر یک از حمله‌های این مجموعه را با بیان نوع و خصوصیات اصلی آن به صورت کوتاه معرفی کنید.

سوال ۴ مقاومت تضمین‌شده[۴] (۲۰ نمره)

فرض کنید برای مسئله دسته‌بندی با برچسب‌های $y \in \{-1, 1\}$ از یک مدل دسته‌بند خطی استفاده می‌کنیم. خروجی این مدل به صورت $f(x) = sign(w^T x + b)$ است که در آن w شامل وزن های مدل است.

(الف) (۱۰ نمره) شعاع تضمین مقاومت [۵] این مدل در برابر اغتشاش جمع‌شونده با ورودی را بیابید.

[۳] Transformation
[۴] Certified Robustness
[۵] Robustness Certified Radius

(ب) (۱۰ نمره) فرض کنید مشابه با روش هموارسازی تصادفی [۶] ورودی‌های x را ابتدا با یک نویز گاوسی به صورت $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ جمع می‌کنیم و سپس به عنوان ورودی به دسته‌بند خطی می‌دهیم. حال تابع زیر را تعریف می‌کنیم:

$$g(x) = \underset{c \in \{0,1\}}{\operatorname{argmax}} P[f(x + \epsilon) = c]$$

که در آن تابع g محتمل‌ترین کلاس برای حالات مختلف خروجی دسته‌بند هنگام جمع با نویز را نشان می‌دهد. نشان دهید همواره داریم $g(x) = f(x)$. به عبارت دیگر نشان دهید هرگاه یک نویز گاوسی با میانگین صفر با ورودی یک دسته‌بند خطی جمع شود، فارغ از پارامتر واریانس آن، همواره محتمل‌ترین کلاس برای خروجی داده‌ی جدید، همان خروجی دسته‌بند به ازای ورودی اصلی است.

سوال ۵ تقطیر دفاعی[۷] (۳۰ نمره)

بخش‌های مشخص شده در دفترچه‌ی Defensive Distillation.ipynb را کامل کنید.

۱. شبکه‌ی teacher را به صورت استاندارد و با Temperature بالا آموزش دهید.

۲. با روش Distillation شبکه‌ی student را از روی teacher آموزش دهید.

۳. حمله‌ی FGSM را بر روی خروجی شبکه‌ی student انجام دهید.

۴. حمله‌ی FGSM را بر روی لایه‌ی logit شبکه‌ی student اعمال کنید و دقت حمله را با قسمت قبل مقایسه نمایید.

موفق باشید

[۶] Randomized Smoothing
[۷] Defensive Distillation