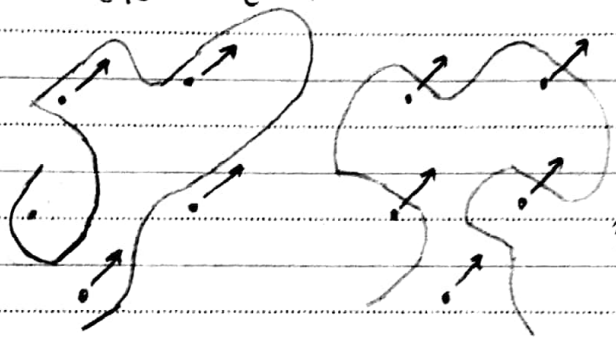
۵
سه‌شنبه
Tuesday
25 Apr 2023
۴ شوال ۱۴۴۴
اردیبهشت

حمیدرضا امیرزاده
۴۰۱۲۰۶۹۹۹

تمرین سری چهارم درس SPML

① الف) آشفتگی های تولید شده در الگوریتم UAP از این لحاظ فراگیر هستند که یک آشفتگی واحد که به درستی توانایی تغییر دادن تمامی داده‌های ورودی را دارد به گونه‌ای که باعث شود تمامی داده هایی که با داده های آشفتگی دیده توزیع یکسانی دارند نیز به داده هایی خصمانه تبدیل شده و مدل آن ها را اشتباه دسته بندی کند.

این آشفتگی ها قابلیت تعمیم کامل به یک مدل متفاوت دیگر را ندارند. زیرا مدل های دیگر دارای مرزهای تصمیم گیری لزوماً یکسانی با مدل اصلی نیستند و همین باعث می‌شود که آشفتگی فراگیر مدل اصلی روی مدل نامذبه به خوبی کار نکند. (زیرا این آشفتگی سعی می‌کند که داده را به فاصله از مرز تصمیم گیری هدایت کند)



به عنوان مثال در شکل راست می بینیم که آشفتگی روی مدل اصلی به خوبی کار می‌کند اما در شکل چپ می بینیم که این آشفتگی روی مدل نامذبه به علت عوض شدن مرزهای تصمیم گیری لزوماً به خوبی عمل نمی‌کند.

ب) وجود آشفتگی خصمانه فراگیر نشان دهنده وجود correlation هندسی بین مرزهای تصمیم گیری در ابعاد بالای است. از آنجایی که آشفتگی به درستی آمده در این روش class-agnostic است بنابراین می‌تواند داده‌های مربوط به هر دسته را به یک دسته غلط تبدیل کند. از اهمیت های کاربردی های UAP می‌توان به موارد زیر اشاره کرد: با شناخت بهتر دلیل وجود این آشفتگی های فراگیر می‌توان به راه حل های بهتر و مقاوم تر برای دفاع شبکه ها در مقابل حملات خصمانه ارائه داد. همچنین می‌توان با ارزیابی مدل های مختلف در مقابل حملات UAP نقاط ضعف مدل را بهتر شناسایی کرد و عملکرد مدل را بهبود داد. از طرف دیگر با بررسی الگوی آشفتگی های حاصل از UAP در مدل ها و دیتاست های مختلف می‌توان به درک بهتری از نحوه عملکرد و تصمیم گیری شبکه های عصبی دست پیدا کرد. از دیگر کاربردهای کلی UAPs می‌توان به یادگیری خصمانه، افزایش مقاومت مدل و همچنین data augmentation

نام برد.

Scanned with CamScanner

ج) v را بردار آشفتگی فراگیر در نظر می‌گیریم. $\hat{k}$ تا تابع دسته‌بند تعریف می‌کنیم. حال یافتن v را این‌گونه با کمک مسئله بهینه‌سازی مقید زیر انجام داد:

$$\underset{v}{\text{argmax}} \quad \mathbb{E}_{x \sim D}[J(x+v)]$$

$$\text{s.t.} \quad \|v\|_p \leq \varepsilon$$

Scanned with CamScanner

چهارشنبه ۲
Wednesday
26 Apr 2023
۵ شوال ۱۴۴۴
اردیبهشت

الف) همان مقاله ذکر شده، بیان نقطه ضعف سیستم های تشخیص انسان در رویت های ظاهری نسبت به حملات خصمانه ایست. نویسندگان روشی برای تولید Patch های کوچک که می تواند روی لباس یا وسایل همراه افراد تعبیه شده و باعث شود که سیستم های تشخیص انسان دچار خطا شوند و آن ها را تشخیص ندهند.

روشی که ابتدا گردیده، شامل آموزش یک شبکه عمیق برای تولید Patch های خصمانه ای که پرگندازی بهینه می کنند که دارای اندازه کوچکی باشند و همچنین با چاپ کردن آن ها در دنیای واقعی اثر روی حدود را از دست ندهند و باعث شود که سیستم های تشخیص انسان دچار خطا شوند.

ب) به عنوان مثال تصاویر دیجیتال نمونه های خصمانه می توانند به راحتی باعث اشتباه مدل شوند اما اگر بخواهیم با پرینت کردن این نمونه ها در دنیای واقعی، ممکن است در اثر فرایند چاپ نویز هایی در تصویر ایجاد شود که باعث گمراهی و یا از بین رفتن اثر خصمانه این نمونه ها شود.

یکی از راه های برطرف کردن اینگونه مشکلات بهینه کردن فرایند ساخت نمونه خصمانه در دنیای واقعی (به عنوان مثال پرینت کردن) به گونه ای است که Feature هایی که ایجاد می شود، از دید دوربین ها و حسگرها ویژگی های خصمانه داشته باشد. به عنوان روش دیگر می توان از حملات فیزیکی مثل ایجاد یک درخشش در مقابل دوربین یا تصویر کردن یک الگوی خاص به همراه نمونه خصمانه دیجیتال استفاده کرد.

ج) یک مثال از عدم تعمیم مناسب نمونه های خصمانه به دنیای فیزیکی مربوط به کار Eykholt et al. 2018 است که می گوید نمونه خصمانه ای که باعث می شود سیستم های تشخیص اشیا متداول تابلوی ایست را تشخیص دهند به دنیای واقعی منتقل کنند که متوجه شد که در اکثر مواقع این نمونه خصمانه نمی تواند در دنیای واقعی باعث اشتباه سیستم ها شود. عوامل این عدم تعمیم نیز می تواند مربوط به شرایط روشنایی محیط، زاویه دید و فاصله تا تابلوی ایست بوده باشد. برای برطرف کردن اینگونه مشکلات می توان آموزش مدل را روی داده های متنوع تری از نظر روشنایی و

محیطی و زاویه دید و سایه ایجنین انجام داد. یک رویکرد دیگر می‌توان اعمال کردن قید های فیزیکی روی بافت نمونه خصمانه باشد. به عنوان مثال می‌توان شرایط روشنایی و محیطی و یا فاصله تا شی را شبیه‌سازی کرد و به عنوان قید به مسئله اضافه کرد.

د) اگر L را برای تابع loss برای اندازه‌گیری تفاوت بین پیش‌بینی دسته‌بند $f(x)$ و برچسب واقعی y باشد و δ برای بیان آشفتگی کوچکی باشد که برای ساخت نمونه خصمانه به داده تمیز x_0 اضافه می‌شود داریم:

تفاوت این مسئله با مسائل استاندارد ساخت نمونه خصمانه این است که در اینجا آشفتگی δ با نمونه‌گیری از یک توزیع از تبدیلات T به دست می‌آید. می‌توان مسئله ساخت نمونه خصمانه به روش EOT را به صورت زیر بیان کرد:

$$\underset{x'}{\text{argmax}} \; \mathbb{E}_{t\sim T}\left[\log P(y_t \mid t(x'))\right]$$

$$\text{s.b.} \quad \mathbb{E}_{t\sim T}\left[\| t(x') - t(x) \|_p\right] \leq \varepsilon$$

$$x \in [0,1]^d$$

Scanned with CamScanner

الف) با توجه به جدول داده شده به نظر حمله روش PGD و تابع هزینه آن عملکرد بهتری نسبت به obfuscating gradient و gradient masking دارد. علت این امر این است که در حملات black box عملکرد بهتری از روش‌های whitebox که بستگی به گرادیان‌ها هستند داشته باشیم. مثلا یک برآیند آن‌هایی است که روش تک قدمی FGSM از روش‌های iterative عملکرد بهتری نسبت به روش‌های gradient obfuscation دارد. روش‌های CW و PGD 20 عملکرد بهتری نسبت به روش‌هایی که در روش‌های one step و iterative ارزیابی شده‌اند دارند. بنابراین باید علاوه بر اینکه معمولا" این روش‌ها بین حملات ارزیابی شده باید ثبت شده که در حملات unbounded آیا می‌توان به نرخ موفقیت ۰٪ رسید یا خیر. اگر نتوان رسید باید دلیل مبنی بر شک کردن به obfuscation ارائه شده. معیار از این distortion bound نباید نرخ موفقیت حملات بتنی بر گرادیان را از این حد پایین‌تر بیاورد. پس این مقاله باید از روش‌هایی مثل DAPA، EOT و Reparameterization استفاده کرده

ب) Square Attack: حمله‌ای به صورت black box می‌باشد که در هر تکرار، پاره‌ای مربع شکل و تصادفی به نویز تصویر اضافه می‌کند. پس با استفاده از random search، این پاره‌های مربع شکل تصادفی را به گونه‌ای آپدیت می‌کند که score بالاتری دریافت کند. همچنین با استفاده از نرم‌های l_2 و l_∞ حداکثر مقدار نویز را کنترل می‌کند. از نظر گفته شده این روش black box است و از هیچ اطلاعات گرادیانی استفاده نمی‌کند. این روش نسبت به سایر روش‌های مبتنی بر جعبه سیاه، عملکرد بهتری از نظر تعداد query و نرخ موفقیت دارد.

FAB(Find and Bind): روشی شبیه حمله FGSM است با این تفاوت که با استفاده از Binary search اندازه بهینه‌ای برای perturbation جستجو می‌کند. با اینکه در این روش از گرادیان‌ها استفاده می‌کند، باز هم می‌تواند روی مدل‌هایی که با استفاده از gradient masking آموزش دیده‌اند، اثرگذار باشد.

روز جهانی کار و کارگر

Scanned with CamScanner

APGD: در حمله PGD، step size ثابت است و به همین علت در فرآیند الگوریتم ممکن است تابع هزینه نسبت به دیتاست نتواند به خوبی افزایش یابد. در حمله APGD ابتدا اندازه قدم بزرگ است و همچنین تعداد تکرارهای الگوریتم به تعدادی بازه، مثل $W_0, ..., W_n$ شکسته می‌شود که در ابتدای هر بازه چک می‌شود که نسبت به ابتدای بازه قبلی با استفاده از اندازه قدم فعلی، تابع هزینه افزایش مناسبی داشته است یا خیر، و اگر نه اندازه قدم را برای ادامه نصف می‌کنیم. همچنین در هر به‌روزرسانی از نوعی momentum نیز استفاده می‌شود که این عمل منجر به افزایش بیشتر مقدار تابع هزینه می‌شود. همچنین نیاز نیست در حمله اندازه قدم به عنوان هایپرپارامتر تنظیم شود.

لازم به ذکر است که این حمله را با توجه به تابع هزینه‌ای که انتخاب می‌کنیم به ۲ صورت انجام می‌دهیم.

حمله $APGD_{CE}$ که از تابع هزینه cross entropy استفاده می‌کند. این تابع هزینه نسبت به شیفت لاجیت‌های z، invariant می‌باشد اما نسبت به rescaling آن‌ها invariant نیست.

حمله $APGD_{DLR}$ که از تابع هزینه Difference of logit ratio استفاده می‌کند. این تابع هزینه نسبت به هر دوی shift و rescaling متغیر لاجیت‌های z، invariant می‌باشد.

Scanned with CamScanner

(۴) الف) با توجه به اسلاید های درس می دانیم که شعاع تعیین صحت یک بیان به جای $R = \frac{\sigma}{2}\left(\phi^{-1}(\underline{P_A}) - \phi^{-1}(\overline{P_B})\right)$ استفاده می شود که در آن $\underline{P_A} = P_A$ و $\overline{P_B} = P_B$ فرض می کنیم. همچنین می دانیم که در دسته بندی دو تایی $P_A = 1 - P_B$ داریم، پس جای گذاری در R داریم: $R = \sigma\,\phi^{-1}(P_A)$

$$P_A = \mathbb{P}\left(f(x+\varepsilon) = g(x)\right) = \mathbb{P}\left(\operatorname{sign}(w^T(x+\varepsilon)+b) = \operatorname{sign}(w^Tx+b)\right)$$

طبق فرض $g(x) = f(x)$ می دانیم که

$$= \mathbb{P}\left(\operatorname{sign}(w^Tx + \sigma\|w\|z + b) = \operatorname{sign}(w^Tx+b)\right) \qquad z \sim N(0,1)$$

حالت اول: $w^Tx + b > 0$

$$P_A = \mathbb{P}\left(w^Tx + \sigma\|w\|z + b > 0\right) = \mathbb{P}\left(z > \frac{-w^Tx-b}{\sigma\|w\|}\right)$$

$$= \mathbb{P}\left(z < \frac{w^Tx+b}{\sigma\|w\|}\right) = \phi\left(\frac{w^Tx+b}{\sigma\|w\|}\right)$$

حالت سوم: $w^Tx + b < 0$

$$P_A = \mathbb{P}\left(w^Tx + \sigma\|w\|z + b < 0\right) = \mathbb{P}\left(z < \frac{-w^Tx-b}{\sigma\|w\|}\right)$$

$$= \phi\left(\frac{-w^Tx-b}{\sigma\|w\|}\right)$$

بنابراین داریم: $P_A = \phi\left(\frac{|w^Tx+b|}{\sigma\|w\|}\right)$

از آنجایی که $R = \sigma\,\phi^{-1}(P_A)$ بنابراین داریم:

$$R = \frac{|w^Tx+b|}{\|w\|}$$

Scanned with CamScanner

۱۰
یکشنبه
Sunday
30 Apr 2023
۹ شوال ۱۴۴۴
اردیبهشت

ب) اگر برای دسته‌بندی بیز از most probable class استفاده کنیم $g(x)=1$ داریم:

$$g(x)=1 \Leftrightarrow P(f(x+\varepsilon)=1) > \frac{1}{2}$$

$$\Leftrightarrow P(sign(w^T(x+\varepsilon)+b)=1) > \frac{1}{2} \qquad \varepsilon \sim N(0,\sigma^2 I)$$

$$\Leftrightarrow P(w^T x + w^T\varepsilon + b \geq 0) > \frac{1}{2}$$

$$\Leftrightarrow P(\sigma\|w\| z \geq -w^T x - b) > \frac{1}{2} \qquad z \sim N(0,1)$$

$$\Leftrightarrow P\left(z \leq \frac{w^T x+b}{\sigma\|w\|}\right) > \frac{1}{2} \qquad \Leftrightarrow \frac{w^T x+b}{\sigma\|w\|} > 0$$

$$\Leftrightarrow w^T x + b > 0 \Leftrightarrow f(x)=1$$

حال اگر برای دسته‌بندی بیز از most probable class استفاده کنیم یعنی $g(x)=1$ آنگاه مشابه اثبات بالا $g(x)=-1=f(x)$

بنابراین می‌توان نتیجه گرفت که همواره $\forall x : f(x)=g(x)$

روز ملی خلیج فارس

Scanned with CamScanner