

فصل اول

مقدمه

فصل دوم

مفاهیم پایه

به منظور بررسی حمله‌های خصمانه و دفاع در یادگیری عمیق^۱، باید با برخی از تعاریف و اصطلاح‌های مهم و پرکاربرد این حوزه آشنا شویم. این تعاریف به نوعی پایه و اساس موضوع را تشکیل می‌دهند و به درک درست و صحیح مطالب ذکر شده در این پژوهش کمک فراوانی خواهند کرد. همچنین، از این تعاریف به عنوان سرنخی برای پژوهش و مطالعه بیشتر در حوزه مورد بررسی استفاده می‌شود. تعاریف می‌توانند میزان گستردگی مطالب ذکر شده در پژوهش را نیز نشان دهند؛ به طوری که هرچقدر دایره تعاریف گسترده‌تر باشد، گستردگی مطالب هم بیشتر است، اما در مورد عمق پژوهش پیش‌رو نمی‌توانند اطلاعات زیادی را در اختیار بگذارند و نیاز است تا با مطالعه کامل پژوهش به عمق آن پی برد.

۱-۲ تعاریف

در این بخش، سعی می‌شود تا با تعاریف مهم در حوزه مورد بررسی و تعاریفی از سایر شاخه‌های مرتبط با این موضوع سخن گفته شود. بررسی و واکاوی بیشتر در مورد هر یک از موضوع‌ها را در متن پژوهش مورد توجه قرار خواهیم داد، اما بررسی‌های بیشتر و عمیق‌تر را بر عهده خواننده خواهیم گذاشت.

۱-۱-۲ اصطلاح‌های مهم

برخی از اصطلاح‌های پرکاربرد این موضوع به این شرح هستند [۱]:

- نمونه‌ها/تصاویر خصمانه: نمونه تغییریافته یک تصویر که با ایجاد مواردی مانند نویز^۲ در تصویر اولیه، و به کمک سایر الگوریتم‌های^۳ یادگیری ماشین^۴، شبکه عصبی^۵ را فریب می‌دهند.
- اختلال خصمانه: نویزی که به تصویر اضافه می‌شود تا آن را دچار اختلال کرده، و به یک نمونه خصمانه تبدیل کند.
- آموزش خصمانه: استفاده از تصاویر بدون اختلال و تصاویر متخاصم به منظور آموزش شبکه عصبی.
- متخاصم: فرد ایجادکننده نمونه خصمانه و یا خود نمونه خصمانه را می‌گویند.
- حمله‌های جعبه سیاه: شیوه‌ای از حمله‌ها که فرد متخاصم بدون اطلاع از مقادیر متغیرهای شبکه^۶، و تنها با اطلاعات محدودی از مدل مانند معماری و شیوه آموزش شبکه عصبی، قصد ایجاد نمونه‌های متخاصم را دارد.

¹Adversarial attacks and defences in deep learning

²Noise

³Algorithms

⁴Machine learning

⁵Neural Network

⁶Parameters

- حمله‌های جعبه سفید: شیوه‌ای از حمله‌ها که فرد متخاصم اطلاع کامل از متغیرهای شبکه و مدل شبکه دارد.
- آشکارساز: یک سازوکار که فقط نمونه‌های متخاصم را تشخیص و ردیابی می‌کند.
- نرخ فریب: درصد تصاویر خصمانه‌ای که منجر به تغییر در متغیرهای شبکه می‌شوند.
- اختلال‌های نامحسوس: اختلال‌هایی در نمونه‌ها که از دید انسان نامشخص هستند.
- حمله‌های هدفمند: حمله‌هایی که مدل را فریب می‌دهند؛ به‌منظور آنکه برچسب^۷ خاصی را برای نمونه‌های خصمانه پیش‌بینی کند.
- مدل تهدید: انواع حمله‌های پنهانی در نظر گرفته شده توسط یک رویکرد را می‌گویند. به‌عنوان مثال حمله‌های جعبه سیاه از این نوع هستند.
- انتقال‌پذیری: میزان توانایی یک نمونه خصمانه برای تأثیر بر روی مدل‌هایی به‌غیر از مدلی که برای آن ساخته شده بود.

۲-۱-۲ نمادگذاری‌ها و ماتریس‌ها

در این بخش، به بررسی ماتریس‌های^۸ مهم و نمادگذاری‌های مطرح شده در پژوهش می‌پردازیم. این نمادها انتقال موضوع را راحت‌تر و سریع‌تر می‌کنند. بررسی و درک صحیح از این نمادها، منجر به درک مفاهیم ارائه شده نیز می‌شود [۲].

- مجموعه‌داده: یک مجموعه‌داده به‌صورت $\{x_i, y_i\}_{i=1}^N$ تعریف می‌شود که در آن x_i داده‌ای با برچسب y_i می‌باشد، و N اندازه مجموعه‌داده می‌باشد.
- شبکه عصبی: یک شبکه عصبی با $f(\cdot)$ نمایش داده می‌شود که x را به‌عنوان ورودی دریافت می‌کند، و خروجی آن پیش‌بینی از ورودی خواهد بود که با $f(x)$ نمایش داده می‌شود.
- خطای بهینه‌سازی: خطای بهینه‌سازی که خطای تخصصی هم نامیده می‌شود، به‌صورت $J(\theta, x, y)$ نمایش داده می‌شود که θ وزن مدل را نمایش می‌دهد.
- کراس-آنتروپی^۹: برای مسائل طبقه‌بندی داده‌ها، کراس-آنتروپی بین $f(x)$ و y به‌صورت $J(f(x); y)$ نشان داده می‌شود.

⁷Lable

⁸Matrix

⁹Cross-entropy

- نمونه خصمانه: یک نمونه داده x' ، نمونه متخاصم x شمرده می‌شود اگر تحت شرط ماتریس فاصله به داده x نزدیک باشد، و همچنین $f(x') \neq y$.

$$x' : D(x', x) < \eta, f(x) \neq y \quad (1-2)$$

در این رابطه ماتریس فاصله به صورت $D(.,.)$ نشان داده می‌شود، و η هم فاصله مجاز بین x و x' است. فاصله η به نوعی آستانه‌ای برای شناخت نمونه‌های خصمانه است.

- ماتریس فاصله: ماتریس فاصله به منظور تشخیص فاصله بین نمونه متخاصم و نمونه اصلی است، تا بتوان تفکیک را صورت داد. مهم‌ترین ماتریس فاصله بین x و x' ماتریس L_p است که به صورت $L_p = \|x - x'\|_p$ تعریف می‌شود.

$$\|v\|_p = (|v_1|^p + |v_2|^p + \dots + |v_d|^p)^{1/p} \quad (2-2)$$

۲-۲ روش‌های تهدید

در این بخش، به بررسی سه روش تهدید می‌پردازیم. همان‌طور که در شکل ۱-۲ مشاهده می‌شود، تفاوت روش‌های تهدید در سطح دانش حمله‌کننده‌ها نسبت به مدل هدف است. در اینجا منظور ما از مدل شبکه هدف، معماری و متغیرهای شبکه هستند.



شکل ۱-۲: سطح دانش حمله‌کننده‌ها در روش‌های تهدید [۳]

۱-۲-۲ جعبه سیاه

در روش تهدید جعبه سیاه^{۱۰}، حمله‌کننده‌ها ساختار معماری شبکه هدف و مقادیر متغیرهای شبکه را نمی‌دانند، و فقط می‌توانند با الگوریتم‌های یادگیری عمیق در ارتباط باشند. در این شیوه از حمله‌ها، تنها راه نفوذ به مدل، پرس‌وجو کردن^{۱۱} از مدل، به منظور پیش‌بینی الگوریتم یادگیری عمیق در برابر

¹⁰Black Box

¹¹Query

ورودی‌های خاص است. حمله‌کننده‌ها همواره نمونه‌های خصمانه را بر روی یک طبقه‌بندی‌کننده^{۱۲} جایگزین - که توسط جفت‌های داده و پیش‌بینی و دیگر نمونه‌های خصمانه پنهان از ناظر انسانی آموزش دیده شده‌اند - می‌سازند. حمله‌های جعبه سیاه همواره می‌توانند مدل‌هایی را که به صورت طبیعی در برابر حملات تجهیز و آموزش دیده نشده‌اند به خطر بیندازند [۲].

شکل ۲-۲ یکی از راهکارهای جدید را که مبتنی بر جست‌وجوی محلی برای ایجاد یک تقریب عددی بر روی گرادینان شبکه است، نشان می‌دهد. این روش اختلال‌های کوچکی را بادقت زیاد بر روی تصاویر ایجاد می‌کند که منجر به تهدید می‌شود.



شکل ۲-۲: نمونه‌ای از راهکارهای جدید در تهدیدهای جعبه سیاه

۲-۲-۲ جعبه خاکستری

در روش تهدید جعبه خاکستری^{۱۳}، حمله‌کننده‌ها ساختار مدل هدف را می‌دانند، اما در مورد وزن‌ها و مقادیر متغیرهای شبکه اطلاعاتی ندارند. همانند روش تهدید جعبه سیاه، تنها راه نفوذ در این روش، پرس‌وجو کردن و درخواست زدن به منظور پیش‌بینی الگوریتم‌های یادگیری عمیق در قبال ورودی‌های خاص است. برخلاف روش جعبه سیاه، حمله‌کننده نمونه‌های خصمانه خود را روی یک مدل جایگزین با همان معماری مدل اصلی بنا می‌کند، و از آگاهی خود نسبت به معماری مدل هدف استفاده می‌کند. این‌گونه حمله‌ها نسبت به روش جعبه سیاه از اطلاعاتی بیشتری نسبت به مدل هدف برخوردار هستند؛ در نتیجه انتظار می‌رود که کارایی بیشتری داشته باشند [۲].

۳-۲-۲ جعبه سفید

در روش تهدید جعبه سفید^{۱۴}، حمله‌کننده‌ها دانش کامل از مدل شبکه هدف (معماری و مقادیر متغیرها) دارند، در نتیجه، می‌توانند به صورت مستقیم نمونه‌های متخاصم را در مدل هدف ایجاد کنند. الگوریتم‌های

¹²Classifier

¹³Gray Box

¹⁴White Box

دفاعی زیادی بر اساس روش‌های تهدید جعبه سفید طراحی شده‌اند، اما در حال حاضر، بیشتر الگوریتم‌های موجود، در قبال این روش‌های تهدید آسیب‌پذیرند [۲].

۳-۲ خلاصه

در این فصل، ابتدا برخی از اصطلاح‌های مهم در حوزه یادگیری عمیق و حمله‌های خصمانه را مطرح کردیم، و بیان کردیم که این اصطلاح‌ها به‌نوعی اساس واکاوی موضوع می‌باشند. در ادامه نیز به بررسی نمادگذاری‌های مطرح در پژوهش پرداختیم، تا بتوانیم مقصود خود را در قالب عبارات ریاضی نشان دهیم. در انتها، سه روش تهدید را بررسی و مقایسه کردیم، و ذکر شد که تفاوت این سه روش در سطح آگاهی حمله‌کننده‌ها از معماری شبکه مدل می‌باشد.

فصل سوم

حمله‌های خصمانه

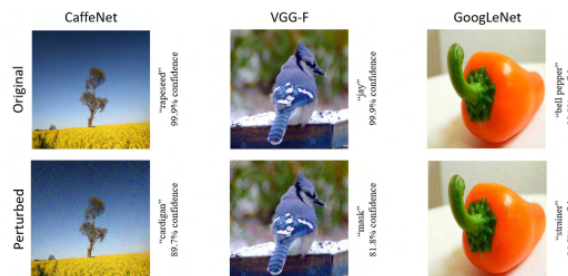
در این فصل، به صورت تخصصی و ریزبینانه به موضوع حمله‌های خصمانه می‌پردازیم، و همچنین حوزه‌هایی را که به واسطه استفاده از یادگیری عمیق دچار تهدید و حمله از سوی افراد متخاصم می‌شوند را بررسی می‌کنیم. در انتهای فصل نیز تلاش می‌کنیم تا تعدادی از الگوریتم‌های حمله که در حال حاضر شناخته شده هستند را معرفی کنیم.

۳-۱ معرفی تعدادی از حوزه‌های مورد حمله

با گسترش استفاده و بهره‌وری از یادگیری عمیق و دیگر زمینه‌های هوش مصنوعی، چالش‌ها و تهدیدها در برابر آن‌ها نیز گسترش پیدا کردند. همچنین، این چالش‌ها وابسته به اینکه از یادگیری عمیق در چه کاربری استفاده می‌شود، ممکن است متفاوت باشند. در ادامه، تعدادی از حوزه‌هایی که از یادگیری عمیق استفاده می‌کنند را بیان می‌کنیم، و چالش‌های آن‌ها را در برابر حمله‌های خصمانه بررسی می‌کنیم.

۳-۱-۱ بینایی کامپیوتر

در یافته‌های [۴] و همکاران، نتایج جالبی در مورد حمله‌های خصمانه به یادگیری عمیق در بینایی ماشین^۱ بیان شده است. به عنوان مثال، علاوه بر اختلال‌های خصمانه در تصویر، وجود اختلال‌های جهانی را در تصاویر نشان می‌دهند که می‌توانند هر نوع طبقه‌بندی از تصاویر را دچار فریب کنند. نمونه‌ای از این اختلال‌های جهانی در شکل ۳-۱ قابل مشاهده است.

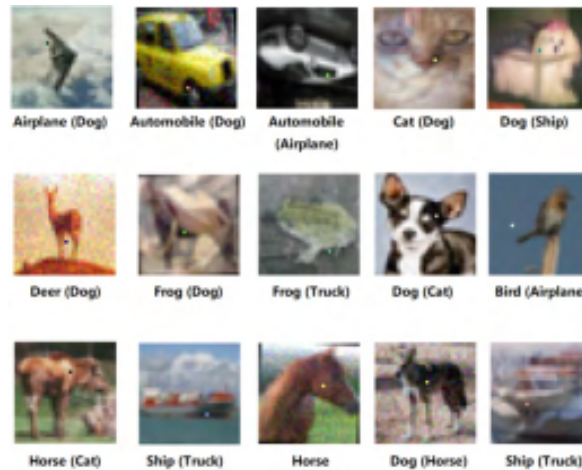


شکل ۳-۱: نمونه‌ای از اختلال‌های جهانی در تصاویر [۲]

به طور مشابه، [۵] و همکاران نشان دادند که حتی چاپ سه‌بعدی اشیاء دنیای حقیقی می‌توانند فرایند یادگیری عمیق در طبقه‌بندی را دچار مشکل کنند. موارد ذکر شده اهمیت بررسی حمله‌های خصمانه در حوزه بینایی ماشین را بیان می‌کنند، در نتیجه، در [۲] سعی شده است تا مطالب مربوط به این موضوع را به صورت جامعی برای خوانندگان بیان کنند، از این رو، دانش کلی از موضوع را می‌توان از آن استخراج کرد. برخی از حمله‌های خصمانه مطرح شده در [۲] به شرح زیر هستند:

^۱Computer vision

۱. حمله‌های تک پیکسلی (One Pixel Attack): یکی از شدیدترین نوع حمله‌ها زمانی است که فقط یک پیکسل در تصویر تغییر کند تا روند طبقه‌بندی را تغییر دهد، و آن را دچار مشکل کند. [۶] مدعی شده است که سه مدل شبکه را با تغییر تنها یک پیکسل با دقت ۷۰.۹۷ درصد فریب دهد. شکل ۲-۳ نمونه‌ای از این حمله‌ها را نشان می‌دهد.



شکل ۲-۳: نمونه‌ای از حمله‌های خصمانه تک پیکسلی [۲]

۲. روش گرادیان سریع (Fast Gradient Sign Method): در پژوهش‌های متعدد، بیان شده است که مقاومت در برابر حمله‌های خصمانه را می‌توان با استفاده از روش‌های یادگیری خصمانه مهار کرد. به‌منظور این کار، [۶] روشی را بیان کرد که بتوان میزان اختلال در تصویر را شناسایی کرد.

۳. دیپ فول (DeepFool): در این روش، کار با یک تصویر تمیز (بدون اختلال) آغاز می‌شود. ابتدا برچسب مربوط به این تصویر را با توجه به طبقه‌بندی که در مدل وجود دارد مشخص می‌کنیم، سپس در هر مرحله، اختلال کوچکی را روی آن اعمال می‌کنیم و برچسب جدید آن را معین می‌کنیم. در انتها تمامی اختلال‌های صورت‌گرفته را جمع می‌کنیم، و به‌عنوان اختلال صورت‌گرفته روی تصویر از ابتدا تا انتها مطرح می‌کنیم. [۶] معتقد است که این روش می‌تواند اختلال‌هایی که کوچک‌تر از روش گرادیان سریع هستند را نیز شناسایی کند.

۲-۱-۳ پردازش زبان طبیعی

پژوهش‌های اخیر نشان می‌دهند که پردازش زبان طبیعی^۲ در برابر حمله‌های خصمانه آسیب‌پذیر است. حمله‌هایی که در سطح کلمات انجام می‌شوند، می‌توانند نمونه‌های خصمانه‌ای با کیفیت بالاتر را تولید کنند، به‌ویژه، در تهدیدهای از نوع جعبه سیاه، این سطح از حمله‌ها کارسازتر هستند. با تمام این

²Natural Language Processing

تفاسیر، حمله‌های موجود نیاز به تعداد زیادی پرس‌وجو برای فریب‌دادن مدل شبکه دارند که در دنیای حقیقی بسیار پرهزینه است؛ از این رو، یافتن روش‌های کم‌هزینه بسیار دشوار و البته ارزشمند است. [۷] و همکارانش، مدل جدیدی را ارائه می‌دهند که از نظر هزینه دارای برتری نسبی در برابر دیگر روش‌ها است. ایده اصلی روش پیشنهاد شده به این صورت است که از مدل‌های خصمانه تولید شده به صورت کامل در مدل شبکه محلی استفاده شود. این کار موجب می‌شود تا شبکه زودتر از موعد آموزش را تکمیل کند، و هزینه بسیار کاهش خواهد یافت.

۲-۳ الگوریتم‌های حمله

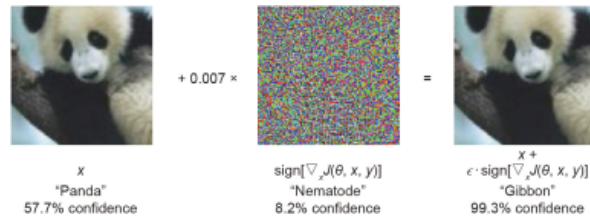
پس از بررسی مقدمات و آشنایی اولیه با موضوع، سعی می‌کنیم تا به صورت عمیق و کاربردی موضوع را پیگیری کنیم. در این بخش قصد داریم تا تعدادی از الگوریتم‌های معروف و پرکاربرد که در حال حاضر مطرح هستند را بررسی کنیم. الگوریتم‌هایی که در زیر مطرح می‌شوند بیشتر در بینایی ماشین و پردازش تصویر مطرح هستند، اما به عنوان پایه‌ای برای الگوریتم‌های استفاده شده در دیگر حوزه‌ها، مورد استفاده قرار می‌گیرند.

۱-۲-۳ ال.بی.اف.جی.اس (L-BFGS)

آسیب‌پذیری شبکه‌های عصبی عمیق در برابر نمونه‌های خصمانه، برای اولین بار در [۸] مطرح شد. در این مقاله، مطرح می‌شود که اختلال‌هایی که با چشم انسان قابل درک نیستند می‌توانند آشفتگی‌هایی را در تصاویر ایجاد کنند که نتایج طبقه‌بندی را تغییر دهند. این الگوریتم، درواقع، بهینه‌سازی عددی مبتنی بر گرادیان غیرخطی می‌باشد که تعداد اختلال‌ها در تصویر را کمینه می‌کند. همچنین، در این الگوریتم از حافظه محدودی استفاده شده است. مهم‌ترین مزیت این الگوریتم این است که در تولید نمونه‌های خصمانه بسیار مؤثر می‌باشد، اما معایبی هم دارد که از جمله آن‌ها می‌توان به زمان بر بودن و حجم بالای محاسبات اشاره کرد، در نتیجه، عملی بودن این روش را زیر سؤال می‌برند [۲].

۲-۲-۳ اف.جی.اس.ام (FGSM)

این شیوه، یک الگوریتم حمله یک‌مرحله‌ای است که در جهت افزایش میزان گرادیان هزینه یعنی $J(\theta, x, y)$ حرکت می‌کند. این امر موجب افزایش هزینه می‌شود. این روش به راحتی قابل گسترش و تعمیم به روش دیگری از الگوریتم‌های حمله به نام اف.جی.اس.ام هدفمند می‌باشد. بدین منظور تنها کاری که نیاز است انجام شود این است که به جای برچسب y از برچسب y' - یعنی برچسب هدفی که می‌خواهیم به آن برسیم - استفاده کنیم. شکل ۳-۳ نمونه معروفی از استفاده این روش برای فریب مدل شبکه را نشان می‌دهد [۲].



شکل ۳-۳: نمونه از حمله به روش اف.جی.اس.ام [۲]

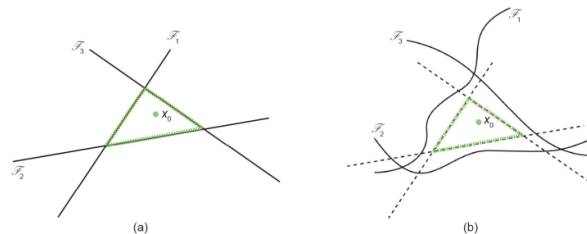
یکی از مزیت‌های مهم این روش نسبت به روش ال.بی.اف.جی.اس این است که زمان محاسباتی بسیار کمتری دارد، در نتیجه، آن را کاربردی‌تر می‌کند.

۳-۲-۳ جی.اس.ام.ای (JSMA)

بر خلاف روش اف.جی.اس.لم، این روش با انتخاب ویژگی‌های مدل، سعی در کم کردن تعداد ویژگی‌های تغییر یافته در هنگام طبقه‌بندی اشتباه را دارد. در این روش، اختلال‌های خطی به صورت مکرر به ویژگی‌های مدل افزوده می‌شوند تا به صورت نامحسوسی مرتبه ویژگی‌ها را افزایش دهند. این روش نسبت به اف.جی.اس.ام تعداد کمتری از ویژگی‌های مدل را تغییر می‌دهد، اما حجم محاسبات بسیار بالاتر خواهد رفت [۲].

۴-۲-۳ دیپ فول (DeepFool)

هدف این روش تولید نمونه‌های خصمانه غیر هدفمند و به حداقل رساندن فاصله اقلیدسی بین نمونه‌های خصمانه و نمونه‌های اصلی است. به طور مکرر در این روش مرزهای بین دسته‌بندی‌ها به روزرسانی می‌شود، و اختلال‌ها به صورت مکرر به تصاویر افزوده می‌شوند. با این تکنیک، می‌توان نمونه‌های خصمانه‌ای با میزان اختلال کمتر، اما قدرت آشفتگی بیشتر در طبقه‌بندی‌ها ایجاد کرد. این روش از لحاظ محاسبات نسبت به جی.اس.ام.ای و اف.جی.اس.ام ضعیف‌تر عمل می‌کند [۲]. شکل ۴-۳ تفاوت سازوکار خطی و غیرخطی در این روش را نشان می‌دهد.



شکل ۴-۳: الف) مدل خطی؛ ب) مدل غیرخطی

۵-۲-۳ گن (GAN)

این روش برای ایجاد نمونه‌های خصمانه در شرایطی که دو شبکه عصبی با هم رقابت می‌کنند، مورد استفاده قرار می‌گیرد. در این شرایط، یکی از شبکه‌ها به‌عنوان تولیدکننده نمونه‌ها عمل کند و دیگری به‌عنوان متمایزکننده نمونه‌ها. این دو شبکه یک بازی برد-باخت را انجام می‌دهند به‌طوری که یک طرف برنده و دیگری بازنده خواهد بود و نه چیزی مابین آنها. در این بازی، تولیدکننده سعی در تولید نمونه‌هایی دارد که شبکه متمایزکننده نتواند آنها را به‌درستی طبقه‌بندی کند، و شبکه متمایزکننده در مقابل، سعی در تفکیک بین نمونه‌های تولید شده توسط شبکه مولد و نمونه‌های اصلی را دارد. مزیت این روش این است که نمونه‌هایی متفاوت از نمونه‌های مورد استفاده در آموزش تولید می‌شود، اما همانند دیگر روش‌ها باید به حجم بالای محاسبات مورد نیاز برای این روش توجه کرد [۲].

۳-۳ خلاصه

در این فصل، موضوع حمله‌های خصمانه را به‌صورت تخصصی‌تر باز کردیم. در ابتدای فصل مطرح کردیم که با توجه به اینکه یادگیری عمیق در حوزه‌های متفاوتی مورد استفاده قرار می‌گیرد، در نتیجه تهدیدهای آن در هر حوزه ممکن است گوناگون باشد. سپس با برخی از حوزه‌های مورد حمله و چالش‌های آن آشنا شدیم، و در انتها نیز چند تا از الگوریتم‌های معروف حمله را مطرح و تا حدودی از لحاظ حجم محاسبات و کاربردی بودن مقایسه کردیم. در فصل آینده در مورد الگوریتم‌های دفاعی در برابر این حمله‌ها مطالبی را بیان می‌کنیم.

فصل چهارم

دفاع در برابر حمله‌های خصمانه

در فصل قبل در مورد حمله‌های خصمانه مواردی را مطرح کردیم؛ در این فصل موضوعی به نام دفاع را در مقابل موضوع حمله بررسی می‌کنیم، و انواع و الگوریتم‌های آن را واکاوی می‌کنیم. در بحث دفاع در برابر حمله‌ها، می‌بایست ابتدا حملات را شناسایی کنیم، و الگوریتم‌های دفاع خود را بر اساس آنها اتخاذ کنیم. بر اساس سطح آگاهی حمله‌کننده‌ها نسبت به مدل شبکه و سرعت تکنیک‌های حمله، باید تکنیک دفاعی خود را انتخاب کنیم. در ادامه به دسته‌بندی کلی تکنیک‌های دفاع می‌پردازیم.

۴-۱ دسته‌بندی تکنیک‌های دفاع

پیش از بررسی الگوریتم‌های دفاع، نیاز است تا با یک دسته‌بندی کلی از تکنیک‌های آن آشنا شویم. تمامی الگوریتم‌هایی که در آینده مطرح خواهند شد در این دودسته قرار می‌گیرند. این دسته‌بندی به ما کمک خواهد کرد تا تصمیم‌گیری خود مبنی بر انتخاب تکنیک‌های دفاعی را، بر اساس تکنیک‌های حمله انجام دهیم.

۴-۱-۱ تکنیک‌های دفاع ابتکاری

دفاع ابتکاری^۱ زمانی مناسب است که بخواهیم در برابر حمله‌های خاصی دفاع انجام دهیم، و تضمینی هم برای میزان دقت دفاع موردنظر نباشد. درواقع، اگر بخواهیم یک سیاست دفاعی در برابر یک حمله خاص منظوره داشته باشیم، از این تکنیک استفاده می‌کنیم. در حال حاضر، موفق‌ترین تکنیک دفاعی ابتکاری، تکنیک آموزش خصمانه^۲ می‌باشد. در تکنیک آموزش خصمانه، نمونه‌های متخاصم را با نمونه‌های دیگر (نمونه‌های آموزش) ترکیب می‌کنیم، و شبکه را با آنها آموزش می‌دهیم. از دسته تکنیک‌های آموزش خصمانه، بهترین دقت را الگوریتم پی.جی.دی^۳ دارا می‌باشد. دسته دیگری از تکنیک‌های ابتکاری نیز وجود دارند که به حذف نویز و تبدیل ورودی‌ها برای کاهش اثر حملات روی داده‌ها وابسته‌اند [۲].

۴-۱-۲ تکنیک‌های دفاع تضمین شده

دفاع‌های تضمین شده^۴، تضمین می‌کنند که هیچ حمله‌ای نمی‌تواند با دقتی فراتر از کران بالای تخمین زده شده رخ دهد، اما این نوع از دفاع‌ها دارای دو عیب می‌باشند: اول اینکه در مقیاس بالا و داده‌های بزرگ کارایی ندارند، و دوم اینکه مدل‌هایی که می‌توانند پشتیبانی کنند بسیار کم است. [۹] اولین دفاع

^۱Heuristic

^۲Adversarial training

^۳PGD

^۴Certificated

ابتکاری را ارائه داد که هر دو مشکل گفته شده یعنی مقیاس‌پذیری و پشتیبانی گسترده را برطرف کرده است. [۹] این دفاع را با نام پیکسل دی.پی (PixelDP) مطرح کرده است. در این الگوریتم همچنین از تکنیک‌های رمزنگاری نیز استفاده شده است. در مقایسه تکنیک‌های دفاع ابتکاری و تضمین شده، می‌توان گفت که کارایی تکنیک‌های ابتکاری به‌خصوص روش آموزش خصمانه از کارایی تکنیک‌های تضمین شده بیشتر است.

۲-۴ الگوریتم‌های دفاع

در ادامه، تعدادی از الگوریتم‌های دفاع^۵ را بررسی می‌کنیم.

۱-۲-۴ آموزش خصمانه

آموزش خصمانه یکی از شیوه‌های دفاع است که تلاش می‌کند با آموزش نمونه‌های خصمانه استحکام شبکه عصبی را بهبود بخشد. در قالب فرمول ریاضی می‌توان آن را به‌صورت بازی کمینه-بیشینه مدل کرد:

$$\min_{\theta} \max_{D(x, x') < \eta} J(\theta, x', y) \quad (1-4)$$

در فرمول بالا $J(\theta, x', y)$ مقدار هزینه خصمانه می‌باشد که مقدار ورودی θ همان وزن‌های شبکه و x' مقدار ورودی خصمانه می‌باشد. y نیز برچسب خروجی به‌ازای ورودی اصلی است.

۲-۲-۴ تصادفی سازی

بسیاری از الگوریتم‌های دفاعی از شیوه تصادفی سازی برای کاهش اثر اختلال‌های خصمانه روی ورودی استفاده می‌کنند. این امر به این علت صورت می‌گیرد که شبکه‌های عصبی عمیق در بیشتر مواقع در برابر اختلال‌های تصادفی مقاوم هستند. یک الگوریتم دفاعی مبتنی بر تصادفی سازی، تلاش می‌کند تا با تصادفی سازی اختلال‌ها، اثر اختلال روی داده‌ها را نیز تصادفی کند، و همان‌طور که گفتیم، اختلال‌های تصادفی برای شبکه‌های عصبی عمیق، نگران‌کننده نیست. این شیوه در برابر تهدیدهای نوع جعبه سیاه و جعبه خاکستری درصدی بالایی از مقاومت را ایجاد می‌کند، اما در برابر تهدیدهای نوع جعبه سفید که حمله‌کننده اطلاع کامل از معماری شبکه دارد، نمی‌تواند مناسب باشد، و حتی ممکن است شبکه را به خطر بیندازد [۱۰].

⁵Defence

۳-۲-۴ رفع آشفستگی

یکی از ساده‌ترین روش‌های دفاعی، رفع آشفستگی^۶ است. رفع آشفستگی را می‌توان بر روی داده‌های ورودی و ویژگی‌های شبکه انجام داد. با رفع آشفستگی روی داده‌های ورودی، می‌توان به‌صورت جزئی و یا به‌طور کامل، اختلال‌های روی ورودی‌ها را حذف کرد. اگر رفع آشفستگی را روی ویژگی‌های شبکه انجام دهیم، سعی داریم تا اختلال‌ها را از روی ویژگی‌های سطح بالایی که شبکه عصبی عمیق فراگرفته است حذف کنیم [۱۰].

۴-۲-۴ دفاع‌های اثبات شدنی

تمامی دفاع‌هایی که در بالا ذکر شد از نوع ابتکاری بودند. همان‌طور که قبلاً هم گفته بودیم، دفاع‌های ابتکاری به‌صورت نظری اثبات نشده‌اند، و فقط به‌صورت تجربی اثربخشی آن‌ها تأیید شده است. بدون اثبات حد بالای خطا به‌صورت نظری، دفاع‌های ابتکاری ممکن است با یک تکنیک حمله جدید، به‌صورت کامل کارایی خود را از دست بدهند. با توجه به مطالبی که ذکر شد، دانشمندان تلاش کرده‌اند که روش‌های دفاعی قابل اثباتی را ارائه دهند که بتوانند همواره درصد دقت مشخصی را در برابر یک شیوه خاص حمله، ارائه دهند [۱۰].

۵-۲-۴ دفاع در شبکه‌های عصبی با وزن‌های کم

گروهی از محققان رابطه مستقیمی بین مقاومت شبکه‌های عصبی و کم بودن وزن‌های شبکه پیدا کرده‌اند. در مدل‌های خطی، نشان داده شده است که بهینه‌سازی روی نمونه‌های خصمانه می‌تواند باعث کاهش ناچیز وزن‌های شبکه شود. در مدل‌های غیرخطی هم کم کردن وزن‌های شبکه، می‌تواند مقاومت را در برابر اختلال‌ها بالا ببرد. محققان مشکل‌های ناشی از پراکندگی وزن‌های شبکه را با اجزای تنظیم‌کننده حل کرده‌اند. بدین منظور از منظم‌سازی به شیوه L_1 استفاده شده است [۱۰].

۳-۴ خلاصه

در این فصل، ابتدا انواع دسته‌های دفاعی در برابر حمله‌های خصمانه را مطرح کردیم. در ادامه فصل نیز به الگوریتم‌های دفاعی پرداختیم، و مزایا و معایب آن‌ها را بیان کردیم، و مشخص کردیم که هر یک از این الگوریتم‌ها در کدام یک از دسته‌های کلی تکنیک‌های دفاع قرار می‌گیرند.

^۶Denoising