دانشگاه صنعتی امیرکبیر
(پلی‌تکنیک تهران)
دانشکده مهندسی کامپیوتر

درس روش پژوهش و ارائه
گزارش نهایی نوشتاری

بررسی حمله‌های خصمانه و دفاع در یادگیری عمیق

نگارش
محمدجواد زندیه

استاد راهنما
دکتر رضا صفابخش

استاد مشاور
دکتر محمد رحمتی

اردیبهشت ۱۴۰۱


بسم الله الرحمن الرحيم

تقدیم به پدر بزرگوار و مادر مهربانم

آن دو فرشته‌ای که از خواسته‌هایشان گذشتند، سختی‌ها را به جان خریدند، و خود را سپر بلای مشکلات و ناملایمات کردند تا من به جایگاهی که اکنون در آن ایستاده‌ام برسم.

سپاس‌گزاری

نهال را "باران" باید، تا سیرابش کند از آب حیات و "آفتاب" باید تا بتاباند نیرو را و محکم کند شاخه‌های تازه روییده را؛ بسی شایسته است تقدیر و تشکر از استاد فرهیخته و فرزانه جناب آقای دکتر رضا صفابخش که با نکته‌های دلاویز و گفته‌های بلند، همواره راهنما و راه‌گشای نگارنده در اتمام و اکمال این پژوهش بوده است.

محمدجواد زندیه
اردیبهشت ۱۴۰۱

چکیده

پس از سال ۲۰۱۳ و با گسترش الگوریتم‌های یادگیری عمیق، ماشین‌ها به‌عنوان موجودیتی شکست‌ناپذیر اطلاق می‌شدند. این تصور به‌گونه‌ای بود که هیچ‌گونه اشتباهی در یادگیری آن‌ها قابل‌قبول نبود، هرچند که می‌توانستند با دقت متفاوتی یادگیری را صورت دهند. این تفکر تا حدودی درست بود، و ماشین‌ها توانسته بودند در مواردی از دقت انسان نیز فراتر رفته و دسته‌بندی تصاویر را بهتر از انسان‌ها انجام دهند. این موضوع، با گذشت زمان و گسترش تکنیک‌های جدید به‌مرور کم‌رنگ شد، و مفهوم حمله‌های خصمانه به‌عنوان رقیبی برای گسترش بی حد و اندازه ماشین‌ها مطرح شدند. با طرح این موضوع، دنیای جدیدی از الگوریتم‌ها سرازیر شدند تا دوباره ماشین را به جایگاه خود بازگردانند. در این پژوهش، قصد داریم تا با تعدادی از الگوریتم‌های حمله و تعدادی از تکنیک‌های دفاع در مقابل آنها آشنا شویم، و دیدی کلی از موضوع پیدا کنیم.

واژه‌های کلیدی:

یادگیری عمیق، روش‌های تهدید، الگوریتم‌های حمله، تکنیک‌های دفاع، آموزش خصمانه

فهرست مطالب

عنوان صفحه

فهرست اشکال

شکل صفحه

فصل اول

مقدمه

شبکه‌های عصبی عمیق به‌صورت فزاینده‌ای در بسیاری از وظایف یادگیری ماشین محبوب و موفق شده‌اند. آن‌ها در مسائل مختلف تشخیصی، در حوزه‌های تصویر، نمودار، متن و گفتار با موفقیت قابل‌توجهی به کار گرفته شده‌اند. در حوزه تشخیص تصویر، آن‌ها قادر به تشخیص اشیا با دقتی نزدیک به انسان هستند؛ همچنین در تشخیص گفتار و پردازش زبان طبیعی نیز استفاده می‌شوند [۴].

با توجه به این دستاوردها، تکنیک‌های یادگیری عمیق در وظایف بحرانی - امنیتی [1] به کار برده می‌شوند. به‌عنوان‌مثال، در وسایل نقلیه خودران، شبکه‌های عصبی پیچیده برای تشخیص علائم جاده استفاده می‌شوند. تکنیک یادگیری ماشین مورد استفاده باید بسیار دقیق، پایدار و قابل‌اعتماد باشد. اگر مدل ما نتواند علامت "ایست" را در کنار جاده تشخیص دهد، و وسیله نقلیه به حرکت خود ادامه دهد، وضعیت خطرناکی خواهد بود. به طور مشابه، در سیستم‌های تشخیص تقلب مالی، شرکت‌ها اغلب از شبکه‌های پیچشی گراف استفاده می‌کنند تا تصمیم بگیرند که آیا مشتریانشان قابل‌اعتماد هستند یا خیر. اگر کلاهبردارانی وجود داشته باشند که اطلاعات هویت شخصی خود را پنهان کنند، ضرر زیادی به شرکت وارد می‌کنند؛ بنابراین، مسائل ایمنی شبکه‌های عصبی عمیق به یک نگرانی اصلی تبدیل شده است [۳].

در سال‌های اخیر، پژوهش‌های زیادی نشان داده است مدل‌های یادگیری عمیق در برابر نمونه‌های خصمانه آسیب‌پذیر هستند. نمونه خصمانه را می‌توان به طور رسمی این‌گونه تعریف کرد: «اختلال‌های خصمانه ورودی‌هایی برای مدل‌های یادگیری ماشینی هستند که مهاجم برای ایجاد اختلال در مدل طراحی کرده است». در حوزه طبقه‌بندی تصویر، این نمونه‌های خصمانه تصاویری هستند که عمداً ترکیب شده‌اند و تا حدود زیادی مشابه تصاویر اصلی هستند، اما می‌توانند طبقه‌بندی‌کننده را برای ارائه خروجی‌ها گمراه کنند. برای یک طبقه‌بندی‌کننده تصویر آموزش‌دیده در مجموعه‌داده‌های ام-نیست [2]، تقریباً تمام نمونه‌های عددی می‌توانند توسط یک اختلال نامحسوس که روی تصویر اصلی اضافه شده است مورد حمله قرار گیرند. در همین حال، در سایر حوزه‌های کاربردی شامل نمودارها، متن یا صدا، طرح‌های حمله خصمانه مشابهی نیز وجود دارد تا مدل‌های یادگیری عمیق را دچار اشتباه کنند. برای مثال، اغتشاش تنها چند یال، می‌تواند شبکه‌های عصبی گراف را گمراه کند، و درج اشتباه‌های تایپی در یک جمله می‌تواند طبقه‌بندی متن یا سیستم‌های گفتگو را فریب دهد. در نتیجه، وجود مثال‌های خصمانه در همه زمینه‌های کاربردی، محققان را نسبت به استفاده مستقیم از شبکه‌های عصبی عمیق در وظایف یادگیری ماشینی بر حذر می‌دارد [۵].

برای مقابله با حمله‌های خصمانه، مطالعه‌هایی برای محافظت از شبکه‌های عصبی عمیق انجام شده است. این رویکردها را می‌توان به سه نوع اصلی دسته‌بندی کرد:

۱. پوشاندن گرادیان [3]: ازآنجایی‌که اکثر الگوریتم‌های حمله بر اساس اطلاعات گرادیان طبقه‌بندی‌کننده‌ها هستند، پوشاندن یا مبهم کردن گرادیان‌ها، مکانیسم‌های حمله را دچار اشتباه می‌کنند [۳].

۲. بهینه‌سازی قوی [4]: این مطالعات نشان می‌دهد که چگونه می‌توان یک طبقه‌بندی‌کننده قوی را

[1] safety-criticcal
[2] MNIST
[3] Gradient masking
[4] Robust optimization

آموزش داد که بتواند نمونه‌های خصمانه را به‌درستی طبقه‌بندی کند [۳].

۳. تشخیص دشمن[۵]: در این رویکرد، سعی می‌شود قبل از آنکه نمونه‌ای برای آموزش به شبکه داده شود، از خوش‌خیم بودن یا خصمانه بودن آن اطلاع یابند. می‌توان آن را به‌عنوان روشی برای محافظت در برابر نمونه‌های متخاصم دید. این روش‌ها مقاومت شبکه عصبی عمیق را در برابر نمونه‌های متخاصم بهبود می‌بخشد [۳].

در این پژوهش، ابتدا با تعدادی از اصطلاح‌های مهم و کاربردی در این حوزه آشنا می‌شویم، و سپس به بررسی روش‌های تهدید و الگوریتم‌های حمله می‌پردازیم. در انتها نیز تعدادی از تکنیک‌های دفاع را بررسی کرده، و با نتیجه‌گیری از مطالب ذکر شده، پیشنهادهای خود را در راستای پیشرفت موضوع بیان می‌کنیم.

[5] Adversary detection

فصل دوم

مفاهیم پایه

به‌منظور بررسی حمله‌های خصمانه و دفاع در یادگیری عمیق [1]، باید با برخی از تعاریف و اصطلاح‌های مهم و پرکاربرد این حوزه آشنا شویم. این تعاریف به‌نوعی 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 به‌عنوان 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؛ به‌طوری 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.

۲-۱ 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.

۲-۱-۱ اصطلاح‌های مهم

برخی از اصطلاح‌های پرکاربرد این موضوع به این شرح هستند:

- نمونه‌ها/تصاویر خصمانه: نمونه تغییریافته یک تصویر که با ایجاد مواردی مانند نویز [2] در تصویر اولیه، و به کمک سایر الگوریتم‌های [3] یادگیری ماشین [4]، شبکه عصبی [5] را فریب می‌دهند [۶].
- اختلال خصمانه: نویزی که به تصویر اضافه می‌شود تا آن را دچار اختلال کرده، و به یک نمونه خصمانه تبدیل کند [۶].
- آموزش خصمانه: استفاده از تصاویر بدون اختلال و تصاویر متخاصم به‌منظور آموزش شبکه عصبی [۶].
- متخاصم: فرد ایجادکننده نمونه خصمانه و یا خود نمونه خصمانه را می‌گویند [۵].
- حمله‌های جعبه سیاه: شیوه‌ای از حمله‌ها که فرد متخاصم بدون اطلاع از مقادیر متغیرهای شبکه [6]، و تنها با اطلاعات محدودی از مدل مانند معماری و شیوه آموزش شبکه عصبی، قصد ایجاد نمونه‌های متخاصم را دارد [۲].

[1] Adversarial attacks and defences in deep learning
[2] Noise
[3] Algorithms
[4] Machine learning
[5] Neural Network
[6] Parameters

- حمله‌های جعبه سفید: شیوه‌ای از حمله‌ها که فرد متخاصم اطلاع کامل از متغیرهای شبکه و مدل شبکه دارد[۲]. آشکارساز: یک سازوکار که فقط نمونه‌های متخاصم را تشخیص و ردیابی می‌کند [۷].

- نرخ فریب: درصد تصاویر خصمانه‌ای که منجر به تغییر در متغیرهای شبکه می‌شوند[۷].

- اختلال‌های نامحسوس: اختلال‌هایی در نمونه‌ها که از دید انسان نامشخص هستند[۵].

- حمله‌های هدفمند: حمله‌هایی که مدل را فریب می‌دهند؛ به‌منظور آنکه برچسب[۷] خاصی را برای نمونه‌های خصمانه پیش‌بینی کند[۵].

- مدل تهدید: انواع حمله‌های پنهانی در نظر گرفته شده توسط یک رویکرد را می‌گویند. به‌عنوان‌مثال حمله‌های جعبه سیاه از این نوع هستند[۱].

- انتقال‌پذیری: میزان توانایی یک نمونه خصمانه برای تأثیر بر روی مدل‌هایی به‌غیر از مدلی که برای آن ساخته شده بود[۱].

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.

- 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[۱].

- شبکه عصبی: یک شبکه عصبی با $f(.)$ 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[۱].

- خطای بهینه‌سازی: خطای بهینه‌سازی که خطای تخاصمی هم نامیده می‌شود، به‌صورت $J(\theta, x, y)$ نمایش داده می‌شود که θ وزن مدل را نمایش می‌دهد[۱].

- کراس-آنتروپی[۹]: برای مسائل طبقه‌بندی داده‌ها، کراس-آنتروپی بین $f(x)$ و y به صورت $J(f(x); y)$ نشان داده می‌شود[۱].

[7] Lable
[8] Matrix
[9] Cross-entropy

- نمونه خصمانه: یک نمونه داده x'، نمونه متخاصم x شمرده می‌شود اگر تحت شرط ماتریس فاصله به داده x نزدیک باشد، و همچنین $f(x') \neq y$ [۱].

$$x' : D(x', x) < \eta, f(x) \neq y \quad (۱-۲)$$

دراین‌رابطه ماتریس فاصله به‌صورت $D(.,.)$ نشان داده می‌شود، و η هم فاصله مجاز بین x و x' است. فاصله η به‌نوعی آستانه‌ای برای شناخت نمونه‌های خصمانه است.

- ماتریس فاصله: ماتریس فاصله به‌منظور 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 به‌صورت $L_p = \|x - x'\|_p$ تعریف می‌شود[۱].

$$\|v\|_p = (|v_1|^p + |v_2|^p + \ + |v_d|^p)^{1/p} \quad (۲-۲)$$

۲-۲ روش‌های تهدید

در این بخش، به بررسی سه روش تهدید می‌پردازیم. همان‌طور که در شکل ۲-۱ 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 ۲-۱: سطح دانش حمله‌کننده‌ها در روش‌های تهدید [۱]

۲-۲-۱ جعبه سیاه

در روش تهدید جعبه سیاه[10]، 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[11] از مدل، به‌منظور پیش‌بینی الگوریتم یادگیری عمیق در برابر ورودی‌های خاص است. حمله‌کننده‌ها همواره نمونه‌های خصمانه را بر روی یک طبقه‌بندی‌کننده[12]

[10] Black Box
[11] Query
[12] Classifier

جایگزین - که توسط جفت‌های داده و پیش‌بینی و دیگر نمونه‌های خصمانه پنهان از ناظر انسانی آموزش‌دیده شده‌اند - می‌سازند. حمله‌های جعبه سیاه همواره می‌توانند مدل‌هایی را که به‌صورت طبیعی در برابر حملات تجهیز و آموزش‌دیده نشده‌اند به خطر بیندازند [۱].

شکل ۲-۲ یکی از راهکارهای جدید را که مبتنی بر جست‌وجوی محلی برای ایجاد یک تقریب عددی بر روی گرادیان 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]

۲-۲-۲ 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 به‌منظور 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 می‌کنند. 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].

در روش تهدید جعبه سفید [۱۴]، حمله‌کننده‌ها دانش کامل از مدل شبکه هدف (معماری و مقادیر متغیرها) دارند، درنتیجه، می‌توانند به‌صورت مستقیم نمونه‌های متخاصم را در مدل هدف ایجاد کنند. الگوریتم‌های دفاعی زیادی بر اساس روش‌های تهدید جعبه سفید طراحی شده‌اند، اما در حال حاضر، بیشتر الگوریتم‌های موجود، در قبال این روش‌های تهدید آسیب‌پذیرند [۲].

[13] Gray Box
[14] White Box

۲-۳ 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

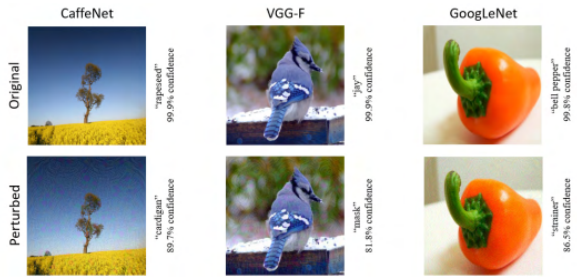
در این فصل، به‌صورت تخصصی و ریزبینانه به موضوع حمله‌های خصمانه می‌پردازیم، و همچنین حوزه‌هایی را که به‌واسطه استفاده از یادگیری عمیق دچار تهدید و حمله از سوی افراد متخاصم می‌شوند را بررسی می‌کنیم. در انتهای فصل نیز تلاش می‌کنیم تا تعدادی از الگوریتم‌های حمله که در حال حاضر شناخته‌شده هستند را معرفی کنیم.

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 [۸] و همکاران، نتایج جالبی در مورد حمله‌های خصمانه به یادگیری عمیق در بینایی ماشین [1] بیان شده است. به‌عنوان‌مثال، 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 قابل‌مشاهده است.



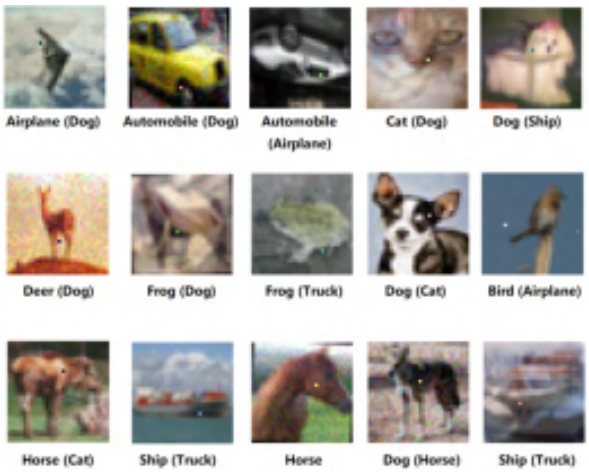
شکل ۳-۱: نمونه‌ای از اختلال‌های جهانی در تصاویر [۱]

به طور مشابه، [۹] 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 [۱] سعی شده است تا مطالب مربوط به این موضوع را به‌صورت جامعی برای خوانندگان بیان کنند، ازاین‌رو، دانش کلی از موضوع را می‌توان از آن استخراج کرد. برخی از حمله‌های خصمانه مطرح شده در [۱] به شرح زیر هستند:

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.

[1] Computer vision

[۱۰] مدعی شده است که سه مدل شبکه را با تغییر تنها یک پیکسل با دقت ۷۰.۹۷ درصد فریب دهد. شکل ۳-۲ نمونه‌ای از این حمله‌ها را نشان می‌دهد.



شکل ۳-۲: نمونه‌ای از حمله‌های خصمانه تک پیکسلی [۱]

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، [۱۰] 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. [۱۰] معتقد است که این روش می‌تواند اختلال‌هایی که کوچک‌تر از روش گرادیان سریع هستند را نیز شناسایی کند.

۳-۱-۲ 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 تفاسیر، حمله‌های موجود نیاز به تعداد زیادی پرس‌وجو برای فریب‌دادن مدل شبکه دارند که در دنیای حقیقی بسیار پرهزینه است؛ ازاین‌رو، یافتن روش‌های کم‌هزینه بسیار دشوار و البته ارزشمند است. [۱۱] و

[2] Natural Language Processing

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 به‌صورت 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.

۳-۲ الگوریتم‌های حمله

پس از بررسی مقدمات و آشنایی اولیه با موضوع، سعی می‌کنیم تا به‌صورت 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 به‌عنوان پایه‌ای برای الگوریتم‌های استفاده شده در دیگر حوزه‌ها، مورداستفاده قرار می‌گیرند.

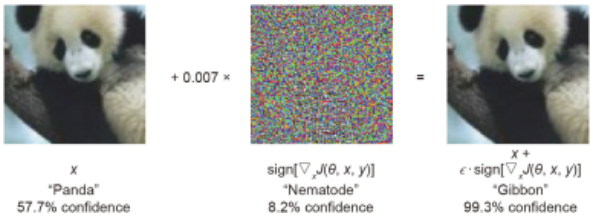
۳-۲-۱ ال.بی.اف.جی.اس (L-BFGS)

آسیب‌پذیری شبکه‌های عصبی عمیق در برابر نمونه‌های خصمانه، برای اولین‌بار در [۵] مطرح شد. در این مقاله، مطرح می‌شود که اختلال‌هایی که با چشم انسان قابل‌درک 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 [۱].

۳-۲-۲ 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 به‌راحتی قابل‌گسترش 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 به‌جای 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 [۱].

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.



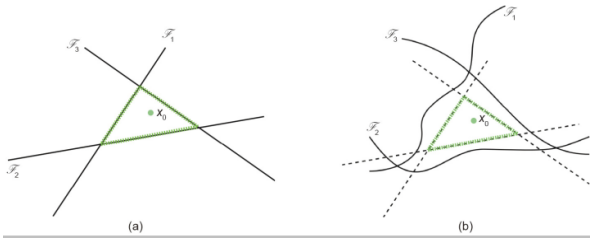
شکل ۳-۳: نمونه از حمله به روش اف.جی.اس.ام [۱]

۳-۲-۳ جی.اس.ام.ای (JSMA)

بر خلاف روش اف.جی.اس.لم، این روش با انتخاب ویژگی‌های مدل، سعی در کم‌کردن تعداد ویژگی‌های تغییریافته در هنگام طبقه‌بندی اشتباه را دارد. در این روش، اختلال‌های خطی به‌صورت مکرر به ویژگی‌های مدل افزوده می‌شوند تا به‌صورت 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 [۱].

۳-۲-۴ دیپ فول (DeepFool)

هدف این روش تولید نمونه‌های خصمانه غیر هدفمند و به‌حداقل‌رساندن فاصله اقلیدسی بین نمونه‌های خصمانه و نمونه‌های اصلی است. به طور مکرر در این روش مرزهای بین دسته‌بندی‌ها به‌روزرسانی می‌شود، و اختلال‌ها به‌صورت 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 [۱]. شکل ۳-۴ تفاوت سازوکار خطی و غیرخطی در این روش را نشان می‌دهد.



شکل ۳-۴: الف) مدل خطی؛ ب) مدل غیرخطی[۳]

۳-۲-۵ گن (GAN)

این روش برای ایجاد نمونه‌های خصمانه در شرایطی که دو شبکه عصبی با هم رقابت می‌کنند، مورداستفاده 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 مورداستفاده در آموزش تولید می‌شود، اما همانند دیگر روش‌ها باید به حجم بالای محاسبات موردنیاز برای این روش توجه کرد [۳].

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 مورداستفاده 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 خاص‌منظوره 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 [۱۲].

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. [۱۳] اولین دفاع ابتکاری را ارائه داد که هر دو مشکل گفته شده یعنی مقیاس‌پذیری و پشتیبانی گسترده را برطرف کرده است. [۱۳] این دفاع را با نام پیکسل دی.پی (PixelDP) مطرح کرده است. در این الگوریتم همچنین

[1] Heuristic
[2] Adversarial training
[3] PGD
[4] Certificated

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.

۴-۲ الگوریتم‌های دفاع

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

۴-۲-۱ 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:

$$min_\theta \ max_{D(x,x')<\eta} J(\theta, x', y) \tag{۴-۱}$$

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.

۴-۲-۲ 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 [۵].

[5] Defence

۴-۲-۳ رفع آشفتگی

یکی از ساده‌ترین روش‌های دفاعی، رفع آشفتگی[۶] است. رفع آشفتگی را می‌توان بر روی داده‌های ورودی و ویژگی‌های شبکه انجام داد. با رفع آشفتگی روی داده‌های ورودی، می‌توان به‌صورت جزئی و یا به طور کامل، اختلال‌های روی ورودی‌ها را حذف کرد. اگر رفع آشفتگی را روی ویژگی‌های شبکه انجام دهیم، سعی داریم تا اختلال‌ها را از روی ویژگی‌های سطح بالایی که شبکه عصبی عمیق فراگرفته است حذف کنیم [۵].

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 [۵].

۴-۲-۵ 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، نشان‌داده‌شده است که بهینه‌سازی روی نمونه‌های خصمانه می‌تواند باعث کاهش ناچیز وزن‌های شبکه شود. در مدل‌های غیرخطی هم کم‌کردن 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 [۵].

۴-۳ 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.

[6] Denoising

فصل پنجم

نتیجه‌گیری

در این پژوهش، پس از بررسی تعدادی از اصطلاح‌های مهم در حوزه حمله‌های خصمانه، به بررسی الگوریتم‌های حمله و همچنین تکنیک‌های دفاع پرداختیم. بیان کردیم که الگوریتم‌های حمله در سه دسته جعبه سفید، جعبه سیاه، و جعبه خاکستری قرار می‌گیرند. در بررسی هر یک از الگوریتم‌های حمله، جایگاه آن‌ها در این دسته‌بندی را نیز بیان کردیم. در ادامه بیان کردیم که تکنیک‌های دفاع در دو دسته ابتکاری و تضمین شده قابل بررسی هستند، و در بررسی الگوریتم‌های دفاع، جایگاه آن‌ها را در این دسته بندی مشخص کردیم. از مهم‌ترین الگوریتم‌های حمله‌ای که بررسی شد، دو حمله تک - خال و اف.جی.اس.ام از اهمیت بالایی برخوردار هستند که شایان توجه بیشتر هستند. در تکنیک‌های دفاعی نیز از کارآمدترین تکنیک‌ها می‌توان به آموزش خصمانه اشاره کرد. در حال حاضر تکنیک‌های دفاعی به‌اندازه‌ای که تکنیک‌های حمله گسترش و دقت پیدا کرده‌اند، پیشرفت نداشته‌اند. به‌عنوان‌مثال، تمامی تکنیک‌های حمله در برابر حمله‌های جعبه سفید آسیب‌پذیرند، و اثر مثبتی را نمی‌توانند نشان دهند [۱].
پس از بیان کلیتی از آنچه که در این پژوهش گذشت، قصد داریم تا تعدادی از پیشنهادها مطرح را بررسی کنیم:

۵-۱ پیشنهادات

در بررسی الگوریتم اف.جی.اس.ام، بیان شد که با اضافه‌شدن نویز به تصویر اصلی می‌توان نمونه خصمانه را تولید کرد. در بهبود و هدفمند کردن این الگوریتم، می‌توان این‌گونه عمل کرد که علاوه بر آنکه نویز ما بتواند شبکه را در دسته‌بندی دچار اشتباه کند، بتواند شبکه را به سمت برچسب موردنظر ما سوق دهد. با این کار، می‌توان امنیت سیستم‌های تشخیص چهره را دچار ضعف کرد، و در مقاصد امنیتی از آن بهره برد.
در پیشنهادی دیگر می‌توان روی بعد بردارها در شیوه حمله‌های تک-خال مطالعاتی را انجام داد که تغییر در مقادیر بردار بتواند شبکه را به سمت برچسب دلخواه سوق دهد، و خوشه خصمانه موردنظر را پدید آورد.
ماهیت کلی این دو پیشنهاد بر این مبنا بود که چگونه علاوه بر ایجاد اختلال، این اختلال‌ها را به صورتی هدفمند برنامه‌ریزی کرد، و علاوه بر هم زدن آموزش صورت گرفته توسط شبکه، شبکه را به شیوه خود آموزش دهیم. در این راستا احتمالا نیاز باشد تا شبکه‌ای را ایجاد کنیم، و نویزها را جهت مقاصد موردنظر آموزش دهیم.

منابع و مراجع

[1] Ren, Kui, Zheng, Tianhang, Qin, Zhan, and Liu, Xue. Adversarial attacks and defenses in deep learning. Engineering, 6(3):346–360, 2020.

[2] Narodytska, Nina and Kasiviswanathan, Shiva Prasad. Simple black-box adversarial attacks on deep neural networks. in CVPR Workshops, vol. 2, 2017.

[3] Xu, Han, Ma, Yao, Liu, Hao-Chen, Deb, Debayan, Liu, Hui, Tang, Ji-Liang, and Jain, Anil K. Adversarial attacks and defenses in images, graphs and text: A review. International Journal of Automation and Computing, 17(2):151–178, 2020.

[4] Papernot, Nicolas, McDaniel, Patrick, Wu, Xi, Jha, Somesh, and Swami, Ananthram. Distillation as a defense to adversarial perturbations against deep neural networks. in 2016 IEEE symposium on security and privacy (SP), pp. 582–597. IEEE, 2016.

[5] Liao, Fangzhou, Liang, Ming, Dong, Yinpeng, Pang, Tianyu, Hu, Xiaolin, and Zhu, Jun. Defense against adversarial attacks using high-level representation guided denoiser. in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1778–1787, 2018.

[6] Akhtar, Naveed and Mian, Ajmal. Threat of adversarial attacks on deep learning in computer vision: A survey. Ieee Access, 6:14410–14430, 2018.

[7] Carlini, Nicholas, Athalye, Anish, Papernot, Nicolas, Brendel, Wieland, Rauber, Jonas, Tsipras, Dimitris, Goodfellow, Ian, Madry, Aleksander, and Kurakin, Alexey. On evaluating adversarial robustness. arXiv preprint arXiv:1902.06705, 2019.

[8] Moosavi-Dezfooli, Seyed-Mohsen, Fawzi, Alhussein, Fawzi, Omar, and Frossard, Pascal. Universal adversarial perturbations. in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1765–1773, 2017.

[9] Athalye, Anish, Engstrom, Logan, Ilyas, Andrew, and Kwok, Kevin. Synthesizing robust adversarial examples. in International conference on machine learning, pp. 284–293. PMLR, 2018.

[10] Moosavi-Dezfooli, Seyed-Mohsen, Fawzi, Alhussein, and Frossard, Pascal. Deepfool: a simple and accurate method to fool deep neural networks. in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2574–2582, 2016.

[11] Zhang, Yu, Yang, Junan, Li, Xiaoshuai, Liu, Hui, and Shao, Kun. Textual adversarial attacking with limited queries. Electronics, 10(21):2671, 2021.

[12] Madry, Aleksander, Makelov, Aleksandar, Schmidt, Ludwig, Tsipras, Dimitris, and Vladu, Adrian. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083, 2017.

[13] Lecuyer, Mathias, Atlidakis, Vaggelis, Geambasu, Roxana, Hsu, Daniel, and Jana, Suman. Certified robustness to adversarial examples with differential privacy. in 2019 IEEE Symposium on Security and Privacy (SP), pp. 656–672. IEEE, 2019.