DOR :20.1001.1.2538161.1401.7.2.2.9

دریافت مقاله: ۱۴۰۱/۰۲/۲۷
پذیرش مقاله:۱۴۰۱/۰۵/۱۸

نوع مقاله: پژوهشی

ارائه روشی جدید و کم هزینه مبتنی بر شبکه عصبی پیچشی برای مقابله با حمله خصمانه بر مدل‌های یادگیری عمیق

علی سرآبادانی
دانشجوی دکتری، دانشکده فنی و مهندسی ـ دانشگاه قم ـ قم ـ ایران
پست الکترونیکی: alisarabadani14@gmail.com

امیر جلالی بیدگلی*
استادیار، دانشکده فنی و مهندسی ـ دانشگاه قم ـ قم ـ ایران
پست الکترونیکی: jalaly@qom.ac.ir

چکیده

روش‌های حمله بـــه مدل‌های یادگیری عمیق می‌توانند برچســـب رده‌هـــا را تغییـــر دهند یـــا این‌کـــه مخاطره‌ای به‌وجـــود آورند و خطرهای جدی امنیتی را ایجاد کنند. در حملات خصمانه مهاجمـــان با ایجاد تغییراتی اندک و البته حساب‌شـــده در داده‌هـــای تصویری، بـــدون این‌که توجه کاربر جلب شود، الگوریتم را به اشتباه می‌اندازند. ازطرفی به واسطه حمله، نوفه‌ها کیفیت تصویر را کاهش می‌دهند و باعث از بین رفتن اطلاعات می‌شوند. نوفهٔ نمک‌وفلفل یکی از محبوب‌ترین نوفه‌هایی است که کیفیت تصویر را تحت تاثیر قرار می‌دهد. روش‌های زیادی برای حذف نوفهٔ نمک‌وفلفل از تصویـــر بـــا حداقل از دســـت دادن اطلاعات پیشـــنهاد شده‌اســـت ولی در این مقاله روشی پیشنهاد شده‌است که نوفهٔ نمک‌وفلفل با روشی مبتنی بر شبکه‌عصبی‌پیچشی در مجموعه داده‌های هدا، CIFAR-10 و MNIST برای مقابله با حملات خصمانه به تصویر اضافه شده‌است. از مزایای این روش پیشنهادی می‌توان به استفاده از کمترین میزان نوفهٔ نمک‌وفلفل،جلوگیری از تعداد بیشتری از حملات خصمانه و همچنین برای مقابله با حملات عمدی در حین حفظ کیفیت تصویر اشـــاره کرد. از نتایج روش پیشنهادی می‌توان به دســـتیابی به درصد موفقیت اســـتحکام در برابر حملات خصمانه، ۸۲/۱۱ در مجموعه داده هدا، ۷۵/۱۹ در مجموعه داده CIFAR-10 و ۷۹/۳٤ در مجموعه داده MNIST اشـــاره کرد که دارای بهبود ۰,۲ در مقایسه با سایر مقالات مروری است. همچنین این مقاله در مقایسه با روشهای پیشین فاقد پیچیدگی محاســـباتی بالا و دارای زمان محاســـباتی کمتر است.

واژه‌های‌کلیدی:جمـــلات خصمانـــه، شـــبکه عصبـــی پیچشی، نوفهٔ نمک‌وفلفل، مدل‌های یادگیری‌عمیق

۱- مقدمه

حمـــلات خصمانه اقدامـــات خرابکارانه‌ای هســـتند که هدفشـــان تضعیف عملکـــرد یادگیری‌عمیـــق، ایجاد رفتار نادرســـت مدل یا کســـب اطلاعات محافظت شـــده اســـت. یادگیری‌عمیق خصمانه در اوایل سال ۲۰۰٤ مورد مطالعه قرارگرفت[۱]. رشد یادگیری‌عمیق و ادغام آن در بسیاری

* نویسندهٔ مسئول

از برنامه‌های کاربردی در سال‌های اخیر باعث تجدید علاقه به یادگیری‌عمیق متخاصم شده‌است.

نگرانی‌های زیادی در حوزه امنیت وجود دارد. می‌توان سیستم‌های مورد حمله را به دستاوردهای هوش مصنوعی مجهز کرد [۲]. برای رده‌بندی تصاویر مدل از کاربر تصاویری دریافت شده است که این تصاویر ممکن است موردحمله متخاصم قرار بگیرد به همین جهت پیش از این بعضی از مدل‌ها برای مقابله با حملات متخاصم اقدام به اضافه کردن نوفهٔ نمک‌وفلفل به تصاویر قبل از پردازش کردن آن‌ها می‌پرداختند که این روش از دو جهت ایراد داشت. در صورت کم یا ناکافی بودن میزان نوفهٔ اضافه شده، مقاومت تصویر در مقابل حملات متخاصم پایین می‌آید. در صورت بیش از اندازه بالا بودن میزان نوفهٔ نمک‌وفلفل کیفیت تصویر و در نتیجه دقت رده‌بندی کاهش می‌یابد. این روش بر مبنای خروجی شبکه عصبی‌پیچشی است. در روش پیشنهادی میزان مناسبی نوفهٔ نمک‌وفلفل در پیکسل‌های مناسب اضافه شده است. با این روش ضمن حفظ کیفیت تصویر، امنیت مدل یادگیری‌عمیق در مقابل حملات خصمانه بالاتر می‌رود. اضافه کردن تعداد زیاد نوفهٔ نمک‌وفلفل باعث کاهش قابل توجه کیفیت تصویر می‌شود. از این رو یافتن میزان مناسب نوفهٔ نمک‌وفلفل و جایگاه پیکسل‌های مناسب قرار گرفتن آن‌ها اهمیت ویژه‌ای دارد. این چالش در روش پیشنهادی بر طرف شده است. میزان و جایگاه پیکسل نوفهٔ نمک‌وفلفل اضافه شده به تصویرها را شبکه‌عصبی‌پیچشی تعیین می‌کند که در این مقاله به‌طور میانگین به تعداد ۸۰ پیکسل است. ساختار مقاله نیز به این صورت است که، در قسمت۲ ادبیات موضوع مقاله مطرح شده‌است.در قسمت ۳ پیشینه پژوهش ارائه گردیده است.در قسمت ٤ روش پیشنهادی قرار گرفته‌است. قسمت ٥ ارزیابی و قسمت ٦ جمع‌بندی می‌باشد. از نتایج روش پیشنهادی می‌توان به دستیابی به درصد موفقیت استحکام در برابر حملات خصمانه، ۸۲/۱۱ در مجموعه داده هدا، ۷۵/۱۹ در مجموعه داده CIFAR-10 و ۷۹/۳٤ در مجموعه داده MNIST اشاره کرد.

۲- ادبیات موضوع

در بخش ۲-۱ توضیحاتی در مورد حملات خصمانه علیه مدل‌های یادگیری‌عمیق، نحوه عمل‌کردشان و همچنین محافظت از سیستم‌های یادگیری‌عمیق در مقابل این حملات داده شده‌است. در بخش ۲-۲ به معرفی نوفهٔ نمک‌وفلفل که در روش پیشنهادی از آن استفاده می‌شود، شده‌است. در بخش ۲-۳ نیز تعاریفی از شبکه‌عصبی‌پیچشی مطرح شده‌است.

۲-۱- حملات خصمانه علیه مدل‌های یادگیری‌عمیق

یادگیری عمیق بخشی جدایی‌ناپذیر از بسیاری از برنامه‌هایی است که ما هر روز استفاده می‌کنیم، از قفل تشخیص چهره در آیفون گرفته تا عملکرد تشخیص صدای الکسا و صافی[1]های اسپم در ایمیل‌های ما که انجام می‌پذیرد. اما فراگیر بودن یادگیری‌عمیق همچنین منجر به حملات خصمانه شده‌است، در این حملات، مهاجمان داده‌ها را دچار آشفتگی[2] می‌کنند و آن‌ها را به‌عنوان ورودی به مدل می‌دهند، سپس از طریق این آشفتگی، ورودی‌های تخاصمی ساخته می‌شود که عملکرد مدل را مختل می‌کنند. بنابراین تغییرات و آشفتگی‌هایی که در داده‌ها ایجاد می‌شوند تا آن‌ها را به نمونه‌های متخاصم تبدیل کند، غیر قابل تشخیص نیستند ولی بسیار به نمونهٔ اولیه شبیه هستند که می‌تواند تشخیص آن‌ها را سخت کند. در سال‌های اخیر، مسئله‌ای تحت عنوان آسیب‌پذیری مدل‌های مبتنی بر یادگیری ماشین مطرح گردیده است که نشان می‌دهد مدل‌های یادگیری در مواجهه با آسیب‌پذیری‌ها از مقاومت بالایی برخوردار نیستند. یکی از معروف‌ترین آسیب‌ها یا به بیان دیگر حملات، تزریق مثال‌های تخاصمی به مدل می‌باشد که در این مورد، شبکه‌های عصبی و به ویژه شبکه‌های عصبی عمیق بیشترین میزان آسیب‌پذیری

1- filter
2- perturbation

را دارند. مثال‌های تخاصمی، از طریق افزوده‌شـــدن اندکی نوفهٔ هدفمند به مثال‌های اصلی تولید می‌شوند، به‌طوری که از منظر کاربر انسانی تغییر محسوسی در داده‌ها مشاهده نمی‌شـــود اما مدل‌های یادگیری ماشـــینی در دسته‌بندی داده‌ها به اشتباه می‌افتند[۳].

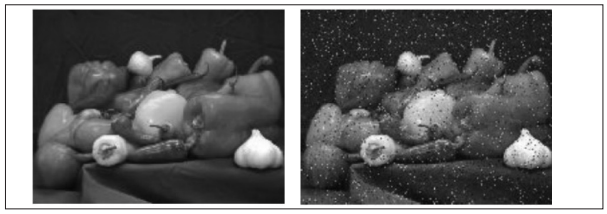
برخلاف سیســـتم‌های نرم‌افزاری کلاسیک، که در آن تولیدکنندگان به‌صورت دستی دستورالعمل‌ها و قوانین را می‌نویســـند، الگوریتم‌های یادگیری‌عمیق رفتار خود را از طریق تجربه توسعه می‌دهند.

به‌عنوان مثال،برای ایجاد یک سیســـتم تشخیص خط، تولیدکنندگان یک الگوریتم‌های یادگیری‌عمیق ایجاد می‌کند و با ارائه تصاویر برچسب‌گذاری‌شده از خطوط خیابان از زوایای مختلف و در شرایط نوری مختلف، آن را آموزش می‌دهد. سپس مدل یادگیری‌عمیق پارامترهای خود را تنظیم می‌کند تا الگوهای رایجی را که در تصاویر حاوی خطوط خیابان رخ می‌دهد، ثبت‌کند. با ســـاختار الگوریتم مناسب و نمونه‌هـــای آموزشـــی کافی، این مدل قـــادر خواهد بود خطوط را در تصاویر و فیلم‌های جدید با دقت قابل توجهی تشـــخیص‌دهد[٤]. اما علیرغم موفقیت آن‌ها در زمینه‌های پیچیده مانند بینایی رایانه و تشـــخیص صدا، الگوریتم‌های یادگیری‌عمیق موتورهای استنتاج آماری هستند، به عبارت دیگـــر توابع پیچیده ریاضی هســـتند کـــه ورودی‌ها را به خروجی تبدیل می‌کنند. اگر یـــک یادگیری‌عمیق تصویری را به‌عنوان حاوی یک شـــیء خاص برچســـب‌گذاری کند، متوجه می‌شـــود که مقادیر پیکســـل در آن تصویر از نظر آماری مشـــابه ســـایر تصاویر شیئی اســـت که در طول آموزش پردازش کرده‌است[٥]. بنابراین با افزودن تکه‌های کوچک و نامحسوس پیکســـل‌ها به یک تصویر، یک عامل مخرب می‌تواند باعث شود الگوریتم یادگیری‌ماشین آن را به‌عنوان یک تصویر نادرست در گروه نادرست رده‌بندی کند. انواع اختلالات اعمال شده در حملات خصمانه به نوع داده هـــدف و اثر مـــورد نظر بســـتگی دارد[٦]. مدل تهدید باید برای روش‌های مختلف داده سفارشی‌ســـازی شود تا به‌طور منطقی خصمانه باشد. به‌عنوان مثال، برای تصاویر و فایل‌هـــای صوتی، منطقی اســـت که اغتشـــاش داده‌های کوچک را به‌عنوان یک مـــدل تهدید درنظر بگیریم، زیرا به راحتی توسط انسان قابل درک نخواهد بود، اما ممکن است مدل هدف را بدرفتار کند و باعث ناهماهنگی بین انســـان و ماشین شود. با این حال، برای برخی از انواع داده‌ها مانند متن، به سادگی با تغییر یک کلمه یا یک نویسه، ممکن است معنایی را مختل کند و به راحتی توســـط انسان شناسایی شـــود. بنابراین، مدل تهدید برای متـــن باید به‌طور طبیعی با تصویر یا صدا متفاوت باشـــد.همچنین این میزان برای تصاویر نیز متفاوت می‌باشد[۷].

۲-۲-نوفهٔ نمک‌وفلفل

هر حسگر دنیای واقعی تحت‌تاثیر درجه خاصی از نوفه قرار می‌گیرد، خواه حرارتی، الکتریکی یا غیره. به‌طورکلی، نوفه می‌تواند توســـط حسگر تصویر و مدار یک اسکنر یا دوربین دیجیتال تولید شود. نوفهٔ تصویر همچنین می‌تواند از دانه‌هـــای فیلم و نوفهٔ غیرقابل اجتناب یک آشکارســـاز فوتون ایده‌آل ایجاد کند. نوفهٔ تصویر یک محصول جانبی نامطلوب ثبت تصویر است که اطلاعات مورد نظر را مبهم می‌کند. معنی اصلی نوفه، ســـیگنال ناخواسته[۳] است. نوفهٔ نمک‌وفلفل نوعی نوفهٔ ضربه‌ای در تصاویر است. ما نوفهٔ نمک‌وفلفل را در نظر می‌گیریم که برای آن مقدار مشخصی از پیکســـل‌های تصویر سیاه یا سفید (نقاط سیاه یا سفید) هســـتند. به‌طور معمول اگر نقاط سیاهی در تصویر وجود داشته باشـــد آن را نوفهٔ فلفل و اگر نقاط سفید در تصویر وجود داشـــته باشد آن را نوفهٔ نمک می‌نامیم که عمدتا در فرآینـــد انتقال اطلاعات رخ می‌دهنـــد. احتمال رخ دادن این نوفه فقـــط در دو مقدار بوده، یا صفر یـــا ۲۵۵ (تصویر ۸ بیتـــی)، یا یک ســـیگنال را نابود می‌کند (صفـــر می‌کند)، یا یک ســـیگنال را کاملا یک می‌کند و چیـــزی بین آن وجود نـــدارد. (مقدار نوفه را با مقدار ســـیگنال جایگزین می‌کند). این نوفه عموماً به دلیل خطا در انتقال داده، خرابی ســـلول

3- Unwanted signal



شکل۱: تاثیر نوفهٔ نمک‌وفلفل بر تصویر[۹]

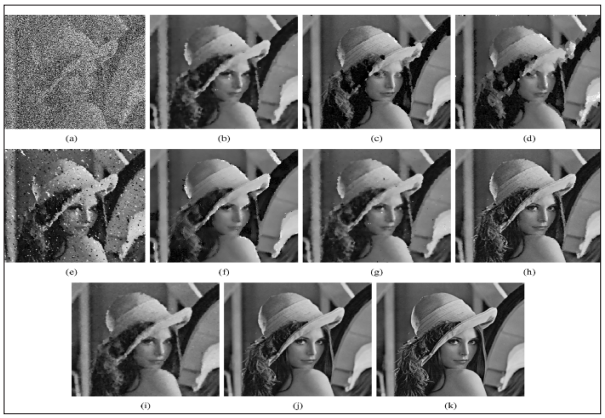
حافظه یا خطاهای مبدل آنالوگ به دیجیتال ایجاد می‌شود. اگر تصویر ۸ بیتی را در نظر گرفته‌شـــود، نوفهٔ نمک‌وفلفل به‌طور تصادفی مقدار معینی از پیکســـل‌ها را در دو حد ۰ یا ۲۵۵ ایجاد می‌کند. نوفه به‌طور قابل توجهی به اطلاعات تصویر آســـیب می‌زند که منجر به مشـــکلاتی در موفقیت در انجـــام وظایف پردازش تصویر مانند تشـــخیص لبه یا تقســـیم‌بندی تصویر می‌شـــود. وظایف تشخیص تصویر از آنجایی که پیکســـل نوفه با اکثر همســـایگان محلی خود متفاوت است، مقدار گرادیان بزرگی مانند پیکسل لبه دارد[۸]. در شـــکل ۱ اعمال نوفـــهٔ نمک‌وفلفل را بر روی تصویر میوه مشاهده می‌کنید:

در شـــکل ۲ تاثیر نوفهٔ نمک‌وفلفل بر تصویر مبتنی بر میزان نوفه اضافه شـــده را مشاهده می‌کنیم. اگر دقت کنید از تصویر k تا تصویر a بـــه مرور میزان نوفهٔ نمک‌وفلفل بیشتر اســـت. در حدی که در تصویر a کیفیت تصویر به شدت کاهش پیدا کرده‌است و پیکسل‌های اصلی تصویر به وسیله نوفه‌های نمک‌وفلفل از بین رفته‌اند. به احتمال فراوان اگر تصویر را بـــه مدلی مبتنی بر یادگیری‌عمیق بدهیم در رده‌بندی تصویر دچار مشکل می‌شود.

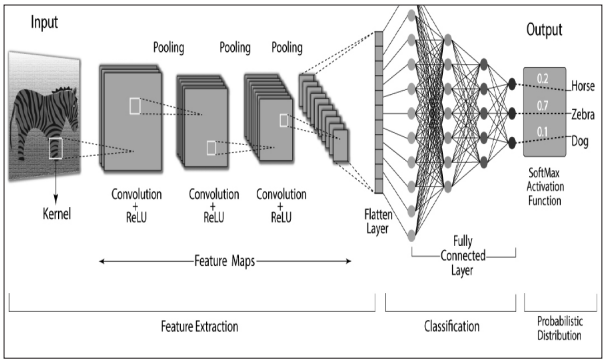
۲-۳- شبکه‌عصبی‌پیچشی

شبکه‌عصبی‌پیچشی که با نام‌های CNN یا ConvNets نیز شـــناخته می‌شـــوند، نوعی شـــبکه عصبی مصنوعی پیش‌خورنده[4] هستند که ساختار اتصال آن از سازمان‌دهی قشـــر بینایی حیوانات الهام گرفته شده‌اســـت. خوشه‌های کوچکی از ســـلول‌ها در قشـــر بینایی به مناطق خاصی از میدان بینایی حســـاس هستند. سلول‌های عصبی منفرد در مغز تنها زمانی واکنش نشـــان می‌دهند که جهت‌گیری‌های

4- feed-forward



شکل۲: تاثیر نوفهٔ نمک‌وفلفل بر تصویر مبتنی بر میزان نوفه اضافه شده[۸]



شکل ۳: معماری شبکهٔ عصبی پیچشی

خاصی از لبه‌ها وجود داشته‌باشـــد. برخـــی از نورون‌ها وقتی لبه‌های عمودی نشان داده می‌شوند، فعال می‌شوند، در حالـــی که برخـــی دیگر هنگامـــی که لبه‌هـــای افقی یا مورب نشـــان داده می‌شوند، واکنش نشـــان می‌دهند[۱۰]. شبکه‌عصبی‌پیچشی نوعی شبکه‌عصبی‌مصنوعی است که در یادگیری‌عمیق برای ارزیابی اطلاعات بصری اســـتفاده می‌شـــود. این شـــبکه‌ها می‌توانند طیف وسیعی از وظایف شـــامل تصاویر، صداها، متون، ویدئوها و سایر رسانه‌ها را انجام دهند. پروفســـور Yann LeCunn از آزمایشـــگاه Bell اولین شبکه‌های‌پیچشی موفق را در اواخر دهه ۱۹۹۰ ایجاد کرد[۱۱]. در چند دهه گذشته، یادگیری‌عمیق به دلیل توانایی آن در مدیریت حجم زیادی از داده‌ها، ابزار بسیار قدرتمندی بوده است. علاقه به استفاده از لایه‌های پنهان[5]، به ویژه در تشـــخیص الگو، از روش‌های ســـنتی پیشـــی گرفته اســـت. یکی از محبوب‌ترین شبکه‌های‌عصبی‌عمیق، شبکه‌های‌عصبی‌پیچشی است.

5- hidden layers

شبکه‌های‌عصبی‌پیچشی دارای یک لایه ورودی، یک لایه خروجی، لایه‌های پنهان متعدد و میلیون‌ها پارامتر هستند که به آن‌ها اجازه می‌دهد اشـیاء و الگوهای پیچیده را یاد بگیرند. از فرآیندهای پیچشـی و ادغام برای نمونه‌برداری فرعـی ورودی داده شـده قبل از اعمال تابع فعال‌سـازی استفاده می‌کند، جایی که همهٔ آن‌ها لایه‌های پنهانی هستند که تا حدی به هم متصل هستند و لایه کاملاً متصل در انتها منجر به لایه خروجی می‌شود. شکل خروجی شبیه به اندازه تصویر ورودی است. پیچش[6] فرآیند ترکیب دو تابع برای تولید خروجی تابع دیگر است. تصویر ورودی با استفاده از صافی‌ها در شبکه‌های‌عصبی‌پیچشـی پیچیده می‌شود که منجر به یک نقشـه ویژگی می‌شود. صافی‌ها وزن‌ها و سوگیری‌هایی هسـتند که به صورت تصادفی بردارهایی در شبکه تولید می‌شوند. شبکه‌های‌عصبی‌پیچشی به جای داشتن وزن‌ها و سوگیری‌های جداگانه برای هر نورون، از وزن‌ها و سوگیری‌های یکسان برای همه نورون‌ها استفاده می‌کند. صافی‌های زیادی را می‌توان ایجاد کرد که هر کدام جنبه متفاوتی از ورودی را می‌گیرد. هسـته‌ها[7] نام دیگری برای صافی‌ها هستند[۱۲].

۳-پیشینه پژوهش

محافظت از سیسـتم‌های یادگیری‌عمیق در مقابل حملات خصمانه

در برابر حملات متخاصم در چند سال گذشته، محققان هوش‌مصنوعی روش‌هـای مختلفی را توسـعه داده‌اند تا مدل‌های یادگیری‌عمیق را در برابر حملات خصمانه قوی‌تر کنند. شناخته‌شده‌ترین روش دفاعی آموزش خصمانه[8] است که در آن یک توسعه‌دهنده آسیب‌پذیری‌ها را با آموزش مدل یادگیری‌عمیق بر روی نمونه‌های متخاصم اصلاح می‌کند. سایر روش‌های دفاعی شامل تغییر یا بهینه‌سازی ساختار مدل است، مانند افزودن لایه‌های تصادفی و برون‌یابی بین چندین مدل یادگیری‌عمیق برای جلوگیری از سوء استفاده از آسـیب‌پذیری‌های متخاصم هر مدل حملات خصمانه را می‌شـود راهی هوشـمندانه برای انجام آزمایش فشار[9] و اشکال‌زدایی[10] روی مدل‌های یادگیری‌عمیق دانست که بالغ[11] در نظر گرفته‌می‌شوند. اگر شما معتقدید که یک فناوری قبل از تبدیل شدن به یک محصول باید به‌طور کامل آزمایش و اشکال‌زدایی شود، پس حمله خصمانه – به منظور آزمایش قوی و بهبود – یک گام اساسی در خط لوله توسعه فناوری یادگیری‌عمیق خواهد بود. محققـان تعدادی راهکار دفاع از حمله متخاصم را پیشـنهاد کرده‌اند که می‌توان آن‌ها را به سه دسـته اصلی، یعنی اصلاح داده‌ها[12]، اصلاح مدل‌ها[13] و استفاده از ابزارهای کمکی[14] تقسیم کرد. ما به ترتیب آن‌ها را با جزئیات شرح می‌دهیم:

۳-۱- اصلاح داده‌ها

ایـن راهکارها به اصلاح مجموعه داده‌های آموزشـی در مرحلـه آموزش یا تغییر داده‌هـای ورودی در مرحله آزمایش، از جمله آموزش خصمانه، پنهان‌سازی گرادیان، مسـدود کـردن قابلیت انتقـال، فشرده‌سـازی داده‌ها و تصادفی‌سازی داده‌ها اشاره دارد.

۱- آموزش خصمانه

در آموزش خصمانه، نمونه‌های متخاصم به مجموعه داده‌های آموزشی معرفی می‌شـوند تا با استفاده از مدل آموزشـی با نمونه‌های متخاصم قانونی، اسـتحکام مدل هـدف را بهبود بخشـند. سـگدی و همـکاران [۱۳] ابتدا نمونه‌های متخاصم را تزریق کرده و برچسـب‌های آن را اصلاح کرد تا مدل در برابر دشمنان قوی‌تر شود.

۲- پنهان‌سازی گرادیان

پنهان‌سازی گرادیان[15]، یک دفاع طبیعی در برابر حملات مبتنی بر گرادیان و حملات با اسـتفاده از روش سـاخت

6- Convolution
7- Kernel
8- adversarial training
9- Pressure
10- Debugging
11- mature
12- modifying data
13- modifying models
14- using auxiliary tools
15- Gradient Hiding

خصمانه مانند FGSM در [۱٤] ارائه شـده است. این روش اطلاعـات مربوط به گرادیـان مدل را از دشـمنان پنهان می‌کنـد، یعنی اگر یک مدل غیرقابل تمایز باشـد (به‌عنوان مثال، یک درخت تصمیم[16]، یک رده‌بندی‌کننده نزدیک‌ترین همسایه[17]، یا یک جنگل تصادفی[18]، حمله مبتنی بر گرادیان نامعتبر اسـت. با این حـال، با یادگیری مدل جعبه‌سـیاه پیشکار[19] با گرادیان و استفاده از نمونه‌های متخاصم تولید شـده توسط این مدل [۱۵]، می‌توان روش را به راحتی در این مورد فریب داد.

۳- مسدود کردن قابلیت انتقال

در مسـدود کردن قابلیت انتقال[20] از آنجایی که ویژگی انتقال پذیری حتی اگـر رده‌بندی‌کننده‌ها معماری متفاوتی داشـته باشـند یا بر روی مجموعه داده‌های مجزا آموزش دیده باشـند، برقرار اسـت، کلید جلوگیـری از حمله جعبه سیاه، جلوگیری از قابلیت انتقال نمونه‌های متخاصم است. حسـینی و همـکاران [۱٦] یک روش سـه مرحله‌ای NULL Labeling را پیشنهاد کردند تا از نمونه‌های متخاصم از یک شبکه به شـبکه دیگر جلوگیری کند. ایده اصلی آن افزودن یک برچسـب NULL جدید بـه مجموعـه داده و رده‌بندی آن‌ها به برچسـب NULL با آمـوزش رده‌بندی کننده برای مقاومت در برابر حملات متخاصم اسـت. این روش به‌طور کلی شـامل سه مرحله اصلی، یعنی آموزش اولیه رده‌بندی کننده هدف، محاسـبه احتمالات NULL و آموزش خصمانه می‌باشد.

۴- متراکم‌سازی داده‌ها

برای متراکم‌سـازی داده‌هـا[21]، Dziugaite و همکاران [۱۷] دریافتند که روش فشرده‌سـازی JPG می‌تواند دقت تشـخیص مدل شبکه را که ناشـی از اختلال حمله FGSM کاهش یافته است بهبود بخشد. داس و همکاران [۱۸] از یک روش فشرده‌سـازی JPEG مشابه برای مطالعه یک روش دفاعی در برابر حملات FGSM و DeepFool اسـتفاده‌کرد. بـا این حال، این فناوری‌های فشرده‌سـازی تصویر هنوز نمی‌توانند به‌عنوان یک دفاع موثر در برابر حملات قدرتمندتر مانند حمـلات Carlini و Wagner عمـل کنند [۱۹] به‌طور مشابه، روش فشرده‌سازی فناوری فشرده‌سازی نمایشگر که در مبارزه با حملات اختلال جهانی اسـتفاده می‌شـود [۲۰] نیز ناکافی است. بزرگترین محدودیت این روش‌های دفاعی مبتنی بر فشرده‌سـازی داده‌ها این اسـت که مقدار زیاد فشرده‌سـازی منجر به کاهش دقت رده‌بندی تصویر اصلی می‌شود، در حالی که مقدار کمی فشرده‌سازی اغلب برای حذف تأثیر اختلال کافی نیست.

۵- تصادفی سازی داده‌ها

در تصادفی‌سازی داده‌ها[22] زی و همکاران [۲۱] نشان داد که عملیات تغییر انـدازه تصادفی نمونه‌های متخاصم می‌توانـد اثربخشـی نمونه‌های متخاصـم را کاهش دهد. به‌طور مشـابه، افـزودن برخی بافت‌هـای تصادفی[23] به نمونه‌هـای متخاصـم نیز می‌توانـد فریب‌دهـی آن‌ها را به مدل شـبکه کاهش دهد. وانگ و همـکاران [۲۲] از یک ماژول تبدیل داده جدا شده از مدل شبکه برای از بین بردن اختلال احتمالی متخاصم در تصویر استفاده کرد و عملیات گسـترده‌ای را در فرآیند آموزش انجام داد، مانند افزودن برخی پردازش‌های تصادفی گاوسـی، که می‌تواند اندکی استحکام مدل شبکه را بهبود بخشد.

۳-۲- اصلاح مدل‌ها

در روش اصلاح مدل شبکه‌عصبی را اصلاح می‌شود، ماننـد با قاعده‌سـازی، تقطیـر دفاعی، فشـردن ویژگی، شـبکه انقباضـی عمیـق، پوشـش دفاعی و شـبکه‌های تجزیه.

16- decision tree
17- nearest neighbor classifier
18- random forest
19- proxy
20- Blocking the Transferability
21- Data Compression
22- Data Randomization
23- random textures

۱- با قاعده‌سازی

هـدف روش با قاعده‌سـازی[۲۴]، بهبـود توانایی تعمیم مدل هدف بـا افزودن عبارات با قاعده اسـت که به‌عنوان عبارات مجازات[۲۵]شناخته می شوند به تابع هزینه[۲۶] و باعث می‌شـود که مدل دارای سـازگاری خوبی برای مقاومت در برابر حملات به مجموعه داده ناشـناخته در پیش‌بینی باشـد. بیگیو و همکاران [۲۳] از یک روش باقاعده‌سـازی برای محدود کردن آسیب‌پذیری داده‌ها هنگام آموزش مدل SVM استفاده کرد. مقالات [۲۶-۲۴] از روش باقاعده‌سازی برای بهبود اسـتحکام الگوریتم استفاده کردند و به نتایج خوبـی در مقاومـت در برابـر حملات متخاصم دسـت یافتند.

۲- تقطیر دفاعی

برای تقطیـر دفاعی[۲۷]، پاپرنوت و همـکاران [۲۷] یک روش تقطیـر دفاعی بـرای مقاومـت در برابر حملات بر اسـاس فناوری تقطیر پیشـنهاد کرد. در مقاله [۲۸] هدف فناوری تقطیر اصلی فشرده‌سـازی مدل در مقیاس بزرگ بـه مقیاس کوچـک و حفظ دقت اولیه اسـت، در حالی که تقطیر دفاعی مقیاس مدل را تغییر نمی‌دهد و مدلی با سطح خروجی صاف‌تر و حساسیت کمتر به اختلال ایجاد می‌کند تا استحکام مدل را بهبود ببخشد.

۳- فشردن ویژگی

فشـردن ویژگی[۲۸] یک روش بهبود مدل است [۲۹] که ایده اصلی آن کاهش پیچیدگی نمایش داده‌ها اسـت و در نتیجه کاهش تداخل متخاصم به دلیل حساسـیت کم است. دو روش اکتشـافی وجـود دارد، یکی کاهـش عمق رنگ در سـطح پیکسـل، یعنی کدگذاری رنگ با مقادیر کمتر و دیگری استفاده از یک صافی مسطح بر روی تصویر است، به‌عنوان مثـال، ورودی‌های متعدد به یک مقدار نگاشـت می‌شـوند، بنابراین مدل را تحت نوفـه و حملات دفاعی[۲۹] ایمن‌تر می‌کند. اگرچه ایـن روش می‌تواند به‌طور موثری از حمـلات خصمانه جلوگیـری کند، اما دقـت رده بندی نمونه‌های واقعی را نیز کاهش می‌دهد.

۴- شبکه انقباضی‌عمیق

در شـبکه انقباضی‌عمیق[۳۰] گو و همکاران [۳۰] نوعی شبکه فشرده‌سـازی‌عمیق[۳۱] را معرفی کرد که از رمزگذار خـودکار کاهش نوفه برای کاهش نوفه دشـمن اسـتفاده می‌کنـد[۳۱]. بر اسـاس این پدیده،DCN ثابـت کرد که اثر دفاعی خاصی در برابر حملاتی مانند L-BGFS دارد.

۵- پوشش دفاعی

در پوشش دفاعی[۳۲]، گائو و همکاران [۳۲] پیشنهاد کرد که یک لایه پوشـش قبل از پردازش مدل شـبکه رده‌بندی شـده درج شـود. ایـن لایه پوشـش تصاویـر اصلی و نمونه‌های متخاصم مربوطه را آموزش داده و تفاوت‌های بین ایـن تصاویر و ویژگی‌های خروجی لایه مدل شـبکه قبلـی را کدگذاری می‌کند. به‌طور کلی اعتقاد بر این اسـت که مهم‌ترین وزن در لایه اضافی مربوط به حسـاس‌ترین ویژگی در شـبکه اسـت. بنابراین، در رده‌بندی نهایی، این ویژگی‌ها با فشار دادن لایه‌های اضافی با وزن اولیه صفر پوشـانده می‌شوند. به این ترتیب می‌توان از انحراف نتایج رده‌بندی ناشی از نمونه‌های متخاصم محافظت کرد.

۶- شبکه‌های تجزیه

در شبکه‌های تجزیه[۳۳]، سیسه و همکاران [۳۳] شبکه‌ای به نام شبکه‌های تجزیه را به‌عنوان روشی دفاعی در برابر حملات متخاصم پیشـنهاد کرد. این شـبکه با کنترل ثابت Lipschitz شـبکه، قاعده‌سـازی سلسـله مراتبی را اتخاذ می‌کند.

۳-۳-استفاده از ابزارهای کمکی

ایـن رویکرد به اسـتفاده از ابزارهای اضافی به‌عنوان ابزار کمکـی برای مدل شـبکه‌عصبی، از جمله -Defense GAN ،MagNet و HGD اشاره دارد.

24- Regularization
25- penalty terms
26- penalty terms
27- Defensive Distillation
28- Feature Squeezing
29- confrontational attack
30- Deep Contractive Network (DCN)
31- deep compression network
32- Mask Defense
33- Parseval Networks

۱- Defense-GAN

سـمنگوی و همکاران [۳۴-۳۵] سازوکاری را پیشنهاد کرد که هم برای حملات جعبه‌سـفید و هـم برای حملات جعبه‌سـیاه قابل اجراسـت تا کارایی اغتشـاش خصمانه را کاهـش دهد. این روش از قدرت شـبکه متخاصم مولد استفاده می‌کند.

۲- MagNet

منـگ و همـکاران [۳۶] چارچوبـی به نـام MagNet پیشـنهاد کرد کـه خروجـی آخرین لایـه رده‌بندی‌کننده را به‌عنـوان یک جعبه‌سـیاه بدون خوانـدن داده‌های لایه داخلی یا تغییر رده‌بندی‌کننـده می‌خواند. MagNet از یک آشکارسـاز برای شناسایی نمونه‌های قانونی و متخاصم استفاده می‌کند. آشکارسـاز[۳۴] فاصله بین یک نمونه معین تحـت آزمایش و نمونه‌های متنـوع را اندازه‌گیری می‌کند و اگر فاصله از آستانه بیشـتر شود، نمونه را رد می‌کند. همچنین از یک اصلاح‌کننده برای تبدیل نمونه متخاصم از طریق یک رمزگذار خودکار به نمونه قانونی مشابه استفاده می‌کند. در حملات جعبه‌سـفید، عملکـرد MagNet به‌طور قابـل توجهی کاهش می‌یابـد زیرا دشـمنان پارامترهای MagNet را می‌دانند. بنابراین، نویسـنده ایده اسـتفاده از چندیـن رمزگذار خودکار و انتخـاب تصادفی یکی در یک زمان را پیشـنهاد کرد تا پیش‌بینی رمزگذار خودکار مورد استفاده دشمن را دشوار کند.

۳- High-Level Representation Guided Denoiser

برخلاف دسـتگاه اسـتاندارد حذف نوفه مانند تابع از دسـت دادن[۳۵] بازسازی در سطح پیکسل، که دارای مشکل تقویت خطا اسـت، HGD[۳۶] می‌تواند به‌طور موثری بر این مشـکل با اسـتفاده از یک تابع از دسـت دادن غلبه کند تا خروجی‌های مدل هدف را با تصویر تمیز مقایسه کند. لیائو و همکاران [۳۷]، HGD را برای طراحی یک مدل هدف قوی در برابر حملات خصمانه جعبه‌سفید و جعبه‌سیاه معرفی کرد. دیگر اسـتفاده از HGD این اسـت که می‌توان آن را بر روی یک مجموعه داده نسـبتا کوچـک آموزش داد و می‌توانـد برای محافظت از مدل‌هایی غیر از مدلی که آن را هدایت می‌کند، استفاده شود. در این مقاله پژوهشگران۲۵ مقاله را در دامنه موضوعی پژوهش بررسـی کرده و پس از بررسـی روش‌های قبلی به یک جمع بنـدی برای مدل پیشـنهادی خود دست یافته اند. جدول (۱) کارهای پیشین بررسی شده را معرفی می‌کند.

۴- روش پیشنهادی

قابل ذکر اسـت که روش پیشـنهادی، دارای نوآوری در مقابل سـایر مقالات و پژوهش‌ها در موضوع استحکام مـدل یادگیری‌عمیق در برابر حملات خصمانه می‌باشـد. همان‌طور که می‌دانیـد در برابر حملات متخاصم در چند سـال گذشـته، محققان هوش‌مصنوعی روشهای مختلفی را توسـعه داده‌اند تا مدل‌هـای یادگیری‌عمیق را در برابر حملات خصمانـه قوی‌تر کنند. شناخته‌شـده‌ترین روش دفاعی آموزش خصمانه اسـت که در آن یک توسعه‌دهنده آسـیب‌پذیری‌ها را با آموزش مدل یادگیری‌عمیق بر روی نمونه‌های متخاصم اصلاح می‌کند. سایر روشهای دفاعی شـامل تغییر یا بهینه‌سازی سـاختار مدل است. هیچ یک از این روش‌ها مبتنی بـر اضافه کردن نوفهٔ نمک‌وفلفل به تصاویر به وسـیله شبکه‌عصبی‌پیچشـی نبوده است. در مقاله در روش پیشـنهادی، قصد بر آن اسـت که در وهله اول یک شبکه عصبی پیچشی را طبق مراحل زیر آموزش داده شود:

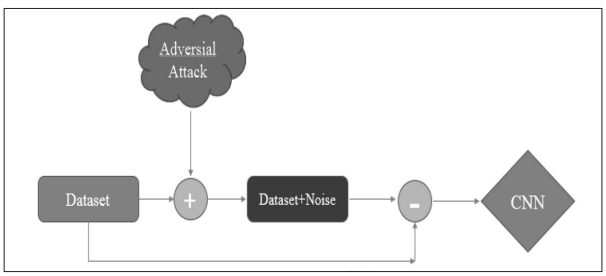
مرحلـه اول- مجموعه داده اولیـه را در معرض حمله خصمانـه قـرار داده و یک مجموعـه داده ثانویه خواهیم داشت.

مرحلـه دوم – مجموعـه داده اولیـه و مجموعه داده ثانویه را با هم مقایسـه کرده و حاصـل اختلاف آن‌ها را به صورت یـک مجموعه داده نهایی بـرای آموزش دادن شبکه‌عصبی‌پیچشـی به کار می‌بریم. می‌دانیم، هر پیکسل عددی معادل ۰ تا ۲۵۵ می‌باشـد ولی عدد دو پیکسل (عدد

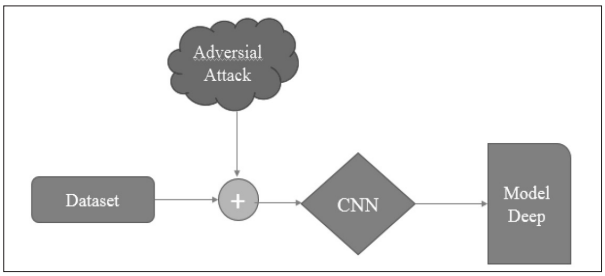
34- detector
35- loss function
36- High-Level Representation Guided Denoiser

جدول ۱: خلاصه مقالات مروری

منابع	نقاط ضعف	نقاط قوت	انواع راهکارها	دسته اصلی
[۱۳]	تزریق نمونه‌های متخاصم	قوی‌تر شدن در برابر دشمنان، استحکام مدل هدف	آموزش خصمانه	اصلاح داده‌ها
[۱۴-۱۵]	فریب سریع	استحکام بیشتر در برابر حملات	پنهان‌سازی گرادیان	
[۱۶]	زمانبر بودن	عدم ورود نمونه‌های متخاصم از یک شبکه به شبکه دیگر	مسدود کردن قابلیت انتقال	
[۱۷-۲۰]	مناسب نبودن در برابر حملات قدرتمند، کاهش دقت رده بندی	روش فشرده‌سازی JPG، افزایش دقت	متراکم‌سازی داده‌ها	
[۲۱-۲۲]	کیفیت بالایی در برابر حملات سخت ندارد.	کاهش اثر بخشی نمونههای متخاصم، کاهش فریب دهی	تصادفی سازی داده‌ها	
[۲۳-۲۶]	نیاز به اعمال قاعده و قوانین دارد.	سازگاری خوب در برابر حملات به مجموعه داده ناشناخته، مقاومت در برابر حملات متخاصم	قاعده‌سازی	اصلاح مدل
[۲۷-۲۸]	مناسب نبودن در برابر حملات قدرتمند	عدم تغییر در مقیاس مدل، حساسیت کمتر به اختلال، استحکام مدل	تقطیر دفاعی	
[۲۹]	کاهش دقت رده بندی نمونه‌های واقعی	کاهش پیچیدگی نمایش داده‌ها، کاهش تداخل متخاصم، ایمن بودن در برابر نوفه	فشردن ویژگی	
[۳۰-۳۱]	عملیات رمزگذاری خود پیچیدگی دارد.	رمزگذار خودکار کاهش نوفه برای کاهش نوفهٔ دشمن	شبکه انقباضی‌عمیق	
[۳۲]	مناسب نبودن در برابر حملات قدرتمند	محافظت از انحراف نتایج رده‌بندی نمونه‌های متخاصم	پوشش دفاعی	
[۳۳]	زمان‌بر بودن پیاده سازی	استفاده از قاعده سازی سلسله مراتبی	شبکه‌های تجزیه	
[۳۴-۳۵]	به نسبت بقیه کارایی خوبی دارد	کاهش کارایی اغتشاش خصمانه	Defense-GAN	ابزارهای کمکی
[۳۶]	کاهش عملکرد در حملات جعبه سفید	استفاده از آشکارساز برای شناسایی نمونه‌های قانونی و متخاصم	Magnet	
[۳۷]	پیچیدگی پیاده سازی	مدل هدف قوی در برابر حملات خصمانه جعبه‌سفید و جعبه‌سیاه، محافظت از مدل‌هایی غیر از مدل مورد استفاده	High-Level Representation Guided Denoiser	



شکل۴: آموزش شبکه‌عصبی‌پیچشی



شکل۵: اضافه کردن نوفهٔ نمک‌وفلفل به وسیله شبکه‌عصبی‌پیچشی و مقابله با حملات خصمانه

منتسب به پیکسل تصویر بدون نوفه، عدد منتسب به پیکسل تصویر با نوفه را از هم کم می‌کنیم عددی بین ۲۵۵ تا ۲۵۵- می‌شود. شکل ۵ این مراحل را نشان می‌دهد.

در وهله دوم، حال که شبکه‌عصبی‌پیچشی ما آموزش داده شده است به‌عنوان ورودی آن مجموعه داده‌ای که مورد حمله خصمانه قرار گرفته است را برای آزمایش به آن می‌دهیم. شبکه‌عصبی‌پیچشی در مرحله آموزش یاد گرفته است که کدام پیکسل‌های تصاویر مورد حمله خصمانه قرار گرفته است. حال با گرفتن ورودی آزمایش، پیکسل‌های مورد حمله را تشخیص داده و در آن پیکسل‌ها

نوفـهٔ نمک‌وفلفـل اضافـه کـرده و خروجـی را به مدل یادگیری‌عمیق می‌دهد. شکل ٦ این مرحله را نشان می‌دهد.

قابل ذکر اسـت که شبکه‌عصبی‌پیچشی ما دارای ٥ لایه می‌باشد و تابع فعال ساز ٤ لایه اولیه ReLU می‌باشد اما تابع فعال سـاز لایه آخر از نوع سیگموئید می‌باشد و همان‌طور کـه می‌دانیم این تابع یک منحنی S شـکل اسـت. زمانی‌که می‌خواهیم خروجی مدل احتمال باشـد، از تابع سیگموئید استفاده می‌کنیم؛ چون تابع سیگموئید مقادیر را به بازه صفر تا ١ می‌برد و احتمالات هم میان همین بازه قرار دارند.

٥-ارزیابی

در این بخش، ابتـدا در بخش ٥-١ به معرفی مجموعه داده‌های استفاده شده در مقاله می‌پردازیم سپس در بخش ٥-٢، ملاحظات و مقایسـه در قالب جدول١ و نمودارهای مختلف به بررسـی خروجی آزمایش‌ها بر روی مجموعه داده‌های معرفی شده در بخش ٥-١ با استفاده از روش‌های نامبرده در بخش روش‌های پیشنهادی می‌پردازیم.

٥-١- مجموعه داده‌های مورد استفاده

در این بخش به معرفی مجموعه داده‌های هدا، CIFAR-10 و MNIST می‌پردازیم.

٥-١-١- مجموعه داده ارقام دست‌نویس هدا

مجموعه ارقام دست‌نویس هدا که اولین مجموعهٔ بزرگ ارقام دست‌نویس فارسی است، مشتمل بر ١٠٢٣٥٣ نمونه دست‌نوشته سیاه سـفید است. خصوصیات این مجموعه داده به شرح زیر است[٣٨]:

درجه تفکیک نمونه‌ها: ٢٠٠ نقطه بر اینچ

تعداد کل نمونه‌ها: ١٠٢٣٥٢ نمونه

تعداد نمونه‌های آموزش: ٦٠٠٠ نمونه از هر رده

تعداد نمونه‌های آزمایش: ٢٠٠٠ نمونه از هر رده

سایر نمونه‌ها: ٢٢٣٥٢ نمونه

در شـکل ٦، نمونه‌هایـی از دسـتخط‌های موجود در مجموعه ارقام دست‌نویس مجموعه داده هدا را نشان داده شده است.

شکل۶: نمونه‌هایی از دستخط های مختلف موجود در مجموعه ارقام دست‌نویس مجموعه داده هدا

٥-١-٢- مجموعه داده CIFAR-10

مجموعه داده CIFAR-10 (انستیتوی تحقیقات پیشرفته کانادا) مجموعه‌ای از تصاویر است که معمولاً برای آموزش الگوریتم‌هـای یادگیری‌ماشـین و بینایی‌رایانه اسـتفاده می‌شود. مجموعه داده CIFAR-10 مجموعه تصاویری است که معمولاً برای آموزش الگوریتم‌های یادگیری‌ماشـین و بینایی رایانه استفاده می‌شـود. این یکی از پرکاربردترین مجموعه‌های داده برای تحقیقات یادگیری‌ماشـین اسـت. مجموعـه داده CIFAR-10 شـامل ٦٠،٠٠٠ تصویـر رنگی ٣٢×٣٢ در ١٠ ردهٔ مختلف اسـت. ١٠ ردهٔ مختلف هواپیما، ماشـین، پرنده، گربه، آهو، سگ، قورباغه، اسب، کشتی و کامیون را نشـان می‌دهد.از هر رده ٦٠٠٠ تصویر وجود دارد. شـبکه‌های عصبی‌پیچشـی بهترین روش تشخیص تصاویر در CIFAR-10 هستند. CIFAR-10 زیرمجموعه‌ای از ٨٠ میلیون عکس که تمامی آن‌ها با اسـتفاده از برچسـب نام‌گذاری شده‌اسـت. در شـکل ٨، نمونه‌هایـی از کیفیت دسـت‌خط‌های موجـود در مجموعه ارقام دسـت‌نویس مجموعه داده هدا را نشان داده شده است.

٥-١-٣- مجموعه داده MNIST

پایـگاه‌داده MNIST[٣٧]، یک پایگاه‌داده بـزرگ از ارقام دسـت‌نویس اسـت که معمولا برای آموزش سیستم‌های مختلف پردازش‌تصویر اسـتفاده می‌شود. این پایگاه‌داده همچنین به‌طور گسـترده‌ای برای آمـوزش و آزمایش در زمینه یادگیری‌ماشین استفاده می‌شود.پایگاه داده MNIST شامل ٦٠٠٠٠ تصویر آموزشی و ١٠٠٠٠ تصویر آزمایشی اسـت. سـازندگان اصلی پایگاه‌داده فهرستی از برخی از

37- Modified National Institute of Standards and Technology database

شکل۷: نمونه‌هایی از تصاویر مختلف موجود در مجموعه مجموعه داده CIFAR-10

روش‌های آزمایش شده بر روی آن را نگه می‌دارند.

۵-۲- ملاحظات و مقایسه

در این بخش نتایج آزمایش‌ها در قالب جدول۱ و چهار نمودار مختلف نشـــان داده و مورد تجزیـــه و تحلیل قرار گرفته‌است.

۵-۲-۱- جدول نتایج

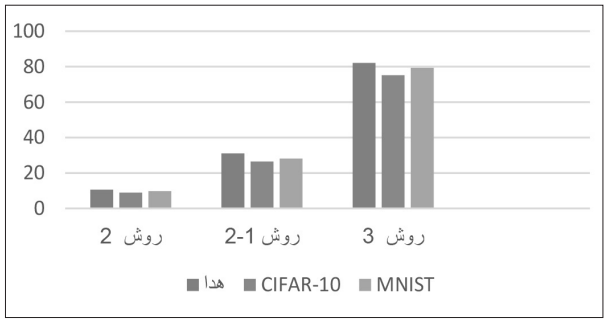
در مجموعـــه داده هـــدا، روش ۱ یعنـــی عـــدم انجام هیچگونـــه تدبیر بـــرای مقابله با حملـــه خصمانه مفاهیم درصد موفقیت، Drop Accuracy و نوفهٔ اضافه شـــده را تعریـــف نمی‌کند. پس ما مقدار عـــددی صفر را در جدول به جهت مقایســـه با سایر مقادیر عددی قرار می‌دهیم. در روش ۲، قبل از ارسال تصاویر به مدل یادگیری‌عمیق برای مقابله با حملات خصمانه بـــه صورت تصادفی مجموعه داده نوفهٔ نمک‌وفلفل را اضافه می‌کنیم در این روش مقدار نوفهٔ اضافه شده در حدود مقدار عددی ۲۰۰ پیکسل و در نتیجه یک مقدار ناکافی نوفه می‌باشـــد که باعث می‌شود درصد موفقیت در حدود مقدار عددی ۱۰/۶۳ باشد. با این حـــال در روش ۲ مقدار Drop Accuracy مقدار کم و قابل قبولی می‌باشـــد. در روش ۲-۱ مجددا نوفهٔ نمک‌وفلفل به جهت مقابله به جهت مقابله با حملات خصمانه به صورت به صورت تصادفی ولی این بار با مقدار نوفهٔ بیشتر یعنی حدود مقدار عددی ۸۰۰ پیکســـل به مجموعه داده اضافه می‌شـــود. در نتیجه، مشـــاهده می‌کنیم بـــه دلیل افزایش

شکل۸: نمونه‌هایی کیفیت از دستخط‌های مختلف موجود در مجموعه ارقام دست‌نویس مجموعه داده MNIST

میزان نوفـــه درصد موفقیت مقابله بـــا حملات خصمانه، افزایش یافته(مقـــدار ۳۱/۰۵) و همزمان مقدار Drop Ac-curacy نیـــز افزایش می‌یابد و به مقدار ۱۲/۱۱ می‌رســـد که تغییر اولی مطلوب و تغییر دومی نامطلوب اســـت. در روش ۳، مشاهده می‌شـــود با استفاده از روش هوشمند اضافه کردن نوفهٔ نمک‌وفلفل توسط شبکه‌عصبی‌پیچشی بـــا مقدار مناســـب و کافی که می‌تواند متغیر باشـــد ولی به‌طور میانگین به میزان ۸۰ می‌باشـــد، به درصد موفقیت بالایی برای مقابله با حملات خصمانه در حین نگه داشتن مقدار عـــددی Drop Accuracy در کـــم ترین حالت خود می‌رسد. در ســـایر مجموعه داده‌ها نیز می‌توان به طریق مشـــابه داده‌های به‌دست آمده را بررسی کرد و با توجه به این‌که مجموعه داده هدا و MNIST تقریبا مشابه یکدیگر هستند انتظار داریم داده‌های جدولشان نیز نزدیک به هم باشند و همچنین به دلیل رنگی بودن تصاویر در مجموعه داده CIFAR-10 توقع داریم که درصد موفقیت برای مقابله با حملات خصمانه،Drop Accuracy در تمامی روش‌ها و در مقایسه با سایر مجموعه داده‌ها کمتر باشد. در روش اصـــلاح داده، این راهکارها به اصـــلاح مجموعه داده‌های آموزشـــی در مرحله آموزش یا تغییـــر داده‌های ورودی در مرحلـــه آزمایش می‌پردازد ولی در روش پیشـــنهادی عملیاتی بر مجموعه داده‌های آموزشی انجام نشده‌است. همچنین در روش اصلاح مدل شبکه عصبی اصلاح می‌شود که در روش پیشـــنهادی هیچگونه تغییر یا جابه‌جایی در لایـــه های مدل یا پارامترهای دیگـــر مدل صورت نگرفته

جدول۱: مقایسه نتایج روش های مختلف در مجموعه داده‌های هدا، CIFAR-10 و MNIST

	مجموعه داده هدا			مجموعه داده CIFAR-10			مجموعه داده MNIST		
	درصد موفقیت	Accuracy Drop	نوفه اضافه شده	درصد موفقیت	Accuracy Drop	نوفه اضافه شده	درصد موفقیت	Accuracy Drop	نوفه اضافه شده
روش ۱	۰	۰	۰	۰	۰	۰	۰	۰	۰
روش ۲	۱۰/۶۳	۵/۱۳	۲۰۰	۸/۹۳	۳/۱۰	۲۰۰	۹/۷۴	۴/۳۳	۲۰۰
روش ۲-۱	۳۱/۰۵	۱۲/۱۱	۸۰۰	۲۶/۴۴	۸/۹۹	۸۰۰	۲۸/۱۵	۱۰/۰۱	۸۰۰
روش ۳	۸۲/۱۱	۳/۲	هوشمندانه(۸۰)	۷۵/۱۹	۲/۴	هوشمندانه(۸۰)	۷۹/۳۴	۲/۷	هوشمندانه(۸۰)



شکل ۹: درصد موفقیت روش‌های مختلف برای هر گروه از مجموعه داده‌ها

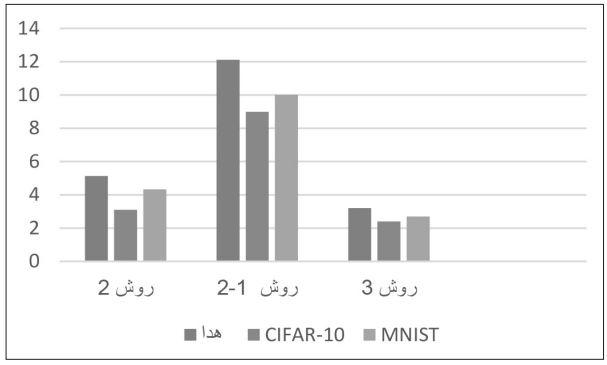
است. در راهکار استفاده از ابزارهای کمکی این رویکرد به استفاده از ابزارهای اضافی به‌عنوان ابزار کمکی برای مدل شبکه‌عصبی، به مدل‌های یادگیری‌عمیق در مقابل حملات خصمانه کمک می‌کند ولی در روش پیشنهادی از هیچ ابزار کمکی استفاده نشده است.

۵-۲-۲- نمودارها

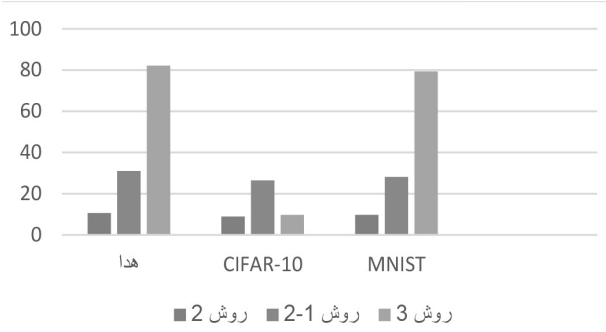
حال سراغ بررسی چهار نمودار صفحه بعدی می‌رویم، که در آن دو نمودار اول به ترتیب درصد موفقیت و Drop Accuracy روش‌های مختلف برای هر گروه از مجموعه داده‌ها می‌باشد. همچنین در نمودارهای سوم و چهارم به ترتیب درصد موفقیت و Drop Accuracy مجموعه داده‌های مختلف برای هر گروه از روش‌های موجود است.

۱- نمودار اول

مشاهده می‌شود که در روش ۳ برای تمامی گروه‌های مجموعه داده بیشترین درصد موفقیت و در روش ۲ کم‌ترین درصد موفقیت را مشاهده می‌شود.



شکل ۱۰: Drop Accuracy روش‌های مختلف برای هر گروه از مجموعه داده‌ها



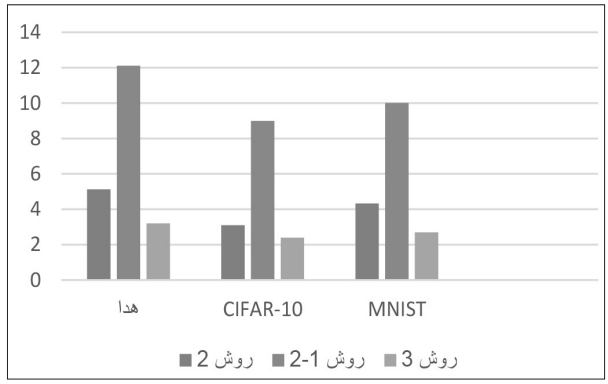
شکل ۱۱: درصد موفقیت مجموعه داده‌های مختلف برای هر گروه از روش‌ها

۲- نمودار دوم

مشاهده می‌شود که به صورت میانگین در روش ۲-۱ برای تمامی گروه های مجموعه داده بیشترین Drop Accuracy و در روش ۳ کم‌ترین Drop Accuracy را مشاهده می‌شود.

۳- نمودار سوم

مشاهده می‌شود که به صورت میانگین در مجموعه داده هدا برای تمامی روش‌ها بیشترین درصد موفقیت را داریم و در مجموعه داده CIFAR-10 کم‌ترین درصد موفقیت را مشاهده می‌شود.



شکل ۱۲: Accuracy Drop مجموعه داده‌های مختلف برای هر گروه از روش‌ها

۴- نمودار چهارم

مشاهده می‌شـــود که به صورت میانگین در مجموعه داده هدا برای تمامی گروه‌های روش‌های موجود بیشترین Drop Accuracy و در مجموعـــه داده CIFAR-10 کم‌تریـــن Drop Accuracy را مشاهده می‌شود.

۶- جمع بندی

همان‌طور که گفته شـــد نوفه به اختلالات ناخواســـته به‌وجود آمده روی تصویر گفته می‌شـــود، به گونه‌ای که بر روی کیفیت تصویر تاثیر منفی می گذارد. یکی از دلایل حذف نوفه از تصویر افزایش کیفیت تصویر جهت نمایش بهتر تصویر می‌باشـــد. از دلایل دیگر حذف نوفه این است که این نوفه مشـــکلی بـــرای یکی از اساســـی‌ترین مراحل پردازش‌تصویر یعنـــی رده بندی می باشـــد. در واقع اگر تصویر نوفه داشته باشد احتمال موفقیت مرحله رده‌بندی کمتر است. در حملات خصمانه مهاجمان با ایجاد تغییرات اندک و البته حساب‌شده در داده‌های تصویری، بدون این‌که توجه کاربر جلب شود، الگوریتم را به اشتباه می‌اندازند و منجر به ایجاد خطرهای جدی امنیتی میشود، برای مقابله با این حملات از روش اضافه کردن نوفۀ نمک‌وفلفل استفاده شـــده است. نوفۀ نمک‌وفلفل یکی از محبوب‌ترین نوفه‌هایی است که کیفیت تصویر را تحت تاثیر قرار می‌دهد. بنابراین در این مقاله روشی پیشنهاد شده که مدعی است که توزیع هوشمند نوفۀ نمک‌وفلفل بر روی تصویر که به‌طور میانگین ۸۰ مورد می‌باشـــد و به وســـیلۀ شبکه‌عصبی‌پیچشـــی‌ای به‌دســـت آمده اســـت، سبب میشـــود که ضمن استفاده از میزان کم و مناسب نوفه کیفیت تصویر حفظ شده و درصد موفقیت بالایی را برای مقابله با حملات متخاصم به همراه داشته باشد. این روش پیشـــنهادی را بر روی ۳ مجموعه داده معروف هدا، CIFAR-10 و MNIST آزمایش شده است و روش خود را با ۳ روش مرسوم دیگر مقایسه شده است و با این مقایســـه قابل درک می‌باشد که روش پیشنهادی، روشی کارآمد و کم هزینه برای مقابله با حملات خصمانه است.

۶- منابع

[1] Cezara Benegui, Radu Tudor Ionescu. Adversarial Attacks on Deep Learning Systems for User Identification based on Motion Sensorsar.arXiv:2009.01109.

[2] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, Debdeep Mukhopadhyay. Adversarial Attacks and Defences: A Survey. https://doi.org/10.48550/arXiv.1810.00069. 28 Sep 2018.

[3] Jianhe Yuan;Zhihai He, "Adversarial Dual Network Learning With Randomized Image Transform for Restoring Attacked Images," IEEE Access· 24 January 2020.

[4] Anirban Chakraborty, Manaar Alam, "Adversarial Attacks and Defences: A Survey ," in https://arxiv.org/abs/1810.00069· 28 Sep 2018.

[5] Gu, S.; Rigazio, L. "Towards deep neural network architectures robust to adversarial examples", 2014. arXiv:1412.5068

[6] Akhtar, N.; Mian, A. "Threat of adversarial attacks on deep learning in computervision: A survey", IEEE Access, vol. 6, pp. 14410-14430, 2018.

[7] Tramèr, F.; Kurakin, A.; Papernot, N.; Goodfellow, I.; Boneh, D.; McDaniel, P. "Ensemble adversarial training: Attacks and defenses, 2017. arXiv:1705.07204.

[8] Ziad Alqadi, "Salt and Pepper Noise: Effects and Removal," in International Journal on Electrical Engineering and Informatics · July 2018.

[9] Raymond Hon Fu Chan, Mila Nikolova, "Salt-and-Pepper Noise Removal by Median-Type Noise Detectors and Detail-Preserving Regularization," in IEEE Transactions on Image Processing · November 2005.

[10] "Convolutional Neural Networks (LeNet) – DeepLearning 0.1 documentation". DeepLearning 0.1. LISA Lab. Retrieved 31 August 2013.

[11] Matusugu, Masakazu; Katsuhiko Mori; Yusuke Mitari; Yuji Kaneda (2003). "Subject independent facial expression recognition with robust face detection using a convolutional neural network" (PDF). Neural Networks. 16 (5): 555–559.

deep neural networks. In Proceedings of the 2016 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22–26 May 2016; pp. 582–597.
[28] Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. arXiv 2015, arXiv:1503.02531.
[29] Xu, W.; Evans, D.; Qi, Y. Feature squeezing: Detecting adversarial examples in deep neural networks. arXiv 2017, arXiv:1704.0115.
[30] Gu, S.; Rigazio, L. Towards deep neural network architectures robust to adversarial examples. arXiv 2014, arXiv:1412.5068.
[31] Rifai, S.; Vincent, P.; Muller, X.; Glorot, X.; Bengio, Y. Contractive auto-encoders: Explicit invariance during feature extraction. In Proceedings of the 28th International Conference on International Conference on Machine Learning, Bellevue, WA, USA, 28 June–2 July 2011; pp. 833–840.
[32] Gao, J.; Wang, B.; Lin, Z.; Xu, W.; Qi, Y. Deepcloak: Masking deep neural network models for robustness against adversarial samples. arXiv 2017, arXiv:1702.06763.
[33] Cisse, M.; Adi, Y.; Neverova, N.; Keshet, J. Houdini: Fooling deep structured prediction models. arXiv 2017, arXiv:1707.05373.
[34] Samangouei, P.; Kabkab, M.; Chellappa, R. Defense-GAN: Protecting classifiers against adversarial attacks using generative models. arXiv 2018, arXiv:1805.06605.
[35] Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Advances in Neural Information Processing Systems, Proceedings of the Annual Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 December 2013; Curran Associates, Inc.: Red Hook, NY, USA, 2014; pp. 2672–2680.
[36] Meng, D.; Chen, H. Magnet: A two-pronged defense against adversarial examples. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, Dallas, TX, USA, 30 October–3 November 2017; pp. 135–147.
[37] Liao, F.; Liang, M.; Dong, Y.; Pang, T.; Zhu, J.; Hu, X. Defense against adversarial attacks using high-level representation guided denoiser. arXiv 2017, arXiv:1712.02976.
[38] Hossein Khosravi , Ehsanollah Kabir , "Introducing a very large dataset of handwritten Farsi digits and a study on their varieties," in Pattern Recognition Letters 28 1133–1141, (2007).
doi:10.1016/S0893-6080(03)00115-1. Retrieved 17 November 2013.
[12] Alzubaidi, L., Zhang, J., Humaidi, A.J. et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 8, 53 (2021). https://doi.org/10.1186/s40537-021-00444-8.
[13] Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. arXiv 2013, arXiv:1312.6199.
[14] Tramèr, F.; Kurakin, A.; Papernot, N.; Goodfellow, I.; Boneh, D.; McDaniel, P. Ensemble adversarial training: Attacks and defenses. arXiv 2017, arXiv:1705.07204.
[15] Papernot, N.; McDaniel, P.; Goodfellow, I.; Jha, S.; Celik, Z.B.; Swami, A. Practical black-box attacks against machine learning. In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, Abu Dhabi, UAE, 2–6 April 2017; pp. 506–519.
[16] Hosseini, H.; Chen, Y.; Kannan, S.; Zhang, B.; Poovendran, R. Blocking transferability of adversarial examples in black-box learning systems. arXiv 2017, arXiv:1703.04318.
[17] Dziugaite, G.K.; Ghahramani, Z.; Roy, D.M. A study of the effect of jpg compression on adversarial images. arXiv 2016, arXiv:1608.00853.
[18] Das, N.; Shanbhogue, M.; Chen, S.T.; Hohman, F.; Chen, L.; Kounavis, M.E.; Chau, D.H. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression. arXiv 2017, arXiv:1705.02900.
[19] Akhtar, N.; Liu, J.; Mian, A. Defense against Universal Adversarial Perturbations. arXiv 2017, arXiv:1711.05929.
[20] Carlini, N.; Wagner, D. Towards evaluating the robustness of neural networks. In Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22–26 May 2017; pp. 39–57.
[21] Xie, C.; Wang, J.; Zhang, Z.; Zhou, Y.; Xie, L.; Yuille, A. Adversarial examples for semantic segmentation and object detection. arXiv 2017, arXiv:1703.08603.
[22] Wang, Q.; Guo, W.; Zhang, K.; Ororbia, I.; Alexander, G.; Xing, X.; Liu, X.; Giles, C.L. Learning adversary-resistant deep neural networks. arXiv 2016, arXiv:1612.01401.
[23] Biggio, B.; Nelson, B.; Laskov, P. Support vector machines under adversarial label noise. In Proceedings of the Asian Conference on Machine Learning, Taoyuan, Taiwan, 13–15 November 2011; pp. 97–112.
[24] Lyu, C.; Huang, K.; Liang, H.N. A unified gradient regularization family for adversarial examples. In Proceedings of the 2015 IEEE International Conference on Data Mining (ICDM), Atlantic City, NJ, USA.14–17 November 2015; pp. 301–309.
[25] Zhao, Q.; Griffin, L.D. Suppressing the unusual: Towards robust cnns using symmetric activation functions. arXiv 2016, arXiv:1603.05145.
[26] Rozsa, A.; Gunther, M.; Boult, T.E. Towards robust deep neural networks with BANG. arXiv 2016, arXiv:1612.00138.
[27] Papernot, N.; McDaniel, P.; Wu, X.; Jha, S.; Swami, A. Distillation as a defense to adversarial perturbations against