



دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

پایان‌نامه‌ی کارشناسی
مهندسی کامپیوتر

عنوان:

کاربرد یادگیری تخصصی در تصاویر پزشکی

نگارش:

سینا کاظمی

استاد راهنما:

دکتر محمدحسین رهبان

۱۴۰۱ تیر

اللهُ أَكْبَرُ

سپاس

از استاد بزرگوارم که با کمک‌ها و راهنمایی‌های بی‌دریغشان، بنده را در انجام این پروژه یاری داده‌اند،
تشکر و قدردانی می‌کنم. همچنین از پدر و مادر عزیزم که همیشه پشتیبانم بودن صمیمانه سپاس‌گزارم.
از جناب آقا حسین یوسفی مقدم تشکر و قدردانی می‌کنم که اجازه دادند بنده از پایان‌نامه ارشد ایشون
استفاده کنم.

چکیده

امروزه هوش مصنوعی در انواع زمینه‌های مختلف از جمله پزشکی نقش برجسته‌ای دارد. با پیشرفت شگرف در بخش سخت‌افزار، استفاده از شبکه‌های عمیق در این علم بیشتر شد. شبکه‌های عمیق قابلیت شناسایی و تخمین توزیع داده را دارند. در زمینه پزشکی چالش اصلی اثر دسته‌ای می‌باشد، اثر دسته‌ای به مختلف شدن توزیع‌های مربوط یک مجموعه داده اشاره دارد که در علم پزشکی کاملاً رایج است، این دسته‌ای شدن می‌تواند عوامل مختلفی همچون تنظیمات دستگاه، ویژگی‌های ظرف‌های مورد آزمایش، یکسان نبودن شرایط محیطی و ... اتفاق بیفتد.

در سال‌های اخیر آسیب‌پذیر بودن مدل‌های یادگیری ماشین به حملات تخاصمی مورد بررسی قرار گرفته است، حملات تخاصمی دسته قابلیت این را دارند که با انحرافی کوچک در داده‌های مجموعه آموزش پیش‌بینی مدل را تغییر دهند (گاهی در راستای میل خود و گاهی به طور تصادفی). یادگیری مقاوم تخاصمی، زمینه‌ای است که به مقابله با حملات تخاصمی می‌پردازد و سعی در افزایش قدرت مدل‌های یادگیری ماشین دارد.

در این پژوهش با استفاده از حمله گرادیان نزول افکنده هدف‌دار و غیر هدف‌دار سعی در افزایش موثر داده‌ها جهت حذف اثر دسته‌ای داریم، در این مقاله با یکسان کردن ویژگی‌های دسته‌ای داده‌ها می‌کوشیم تا مدل‌های یادگیری ماشین را از یادگیری ویژگی‌های دسته‌ای منصرف کنیم و آن‌ها را قادر به یادگیری ویژگی‌های واقعی و مدنظر نماییم. در این پژوهش از دو مجموعه داده در حوزه سیگنال‌های مغزی استفاده می‌شود که دارای اثر دسته‌ای هستند و نشان می‌دهیم که در چهارچوب یادگیری تخاصمی می‌توان دقت مدل‌های یادگیری ماشین را در آزمودن توزیع‌های مختلف چندین درصد افزایش داد.

کلیدواژه‌ها: یادگیری تخاصمی، اثر دسته‌ای، شبکه‌های عمیق، یادگیری مقاوم

فهرست مطالب

۱۱	۱	مقدمه
۱۱	۱-۱	مفاهیم اولیه
۱۱	۱-۱-۱	نمونه‌های تخاصمی
۱۴	۱-۲-۱	یادگیری تخاصمی
۱۵	۱-۳-۱	اهمیت موضوع
۱۵	۱-۴-۱	اثر دسته‌ای
۱۷	۲-۱	طرح مسئله
۱۸	۳-۱	مسئله اسب باهوش
۱۹	۲	بررسی کارهای پیشین
۱۹	۲-۱	عوامل آسیب‌پذیری مدل‌های یادگیری ماشین
۱۹	۲-۱-۱	بیش‌برازش
۲۰	۲-۱-۲	خطی بودن مدل‌های یادگیری ماشین
۲۱	۲-۱-۳	تصمیم‌گیری در نقاط نامعلوم
۲۳	۲-۱-۴	تفاوت توزیع‌ها در زمان آموزش و تست
۲۳	۲-۲	انواع حمله‌های تخاصمی
۲۴	۲-۲-۱	روش علامت گرادیان سریع

۲۵	L-BFGS ۲-۲-۲
۲۵	۳-۲-۲ حمله یک گامی هدف دار
۲۶	۴-۲-۲ روش تکرارشونده ساده
۲۶	۵-۲-۲ نزول گرادیان افکنده شده
۲۷	۶-۲-۲ نقشه بر جستگی مبتنی بر ماتریس ژاکوبین
۲۸	Deep-Fool ۷-۲-۲
۲۸	۸-۲-۲ حمله های جعبه سیاه
۲۸	۹-۲-۲ حمله مرز تصمیم
۲۹	۱۰-۲-۲ احملات همگانی
۳۰	۱۱-۲-۲ الگوریتم های یادگیری تخصصی
۳۰	۱۲-۳-۲ ادبیات موضوع
۳۱	۱۳-۲-۲ یادگیری تخصصی
۳۲	۱۴-۳-۲ تقطیر دفاعی
۳۳	۱۵-۳-۲ یادگیری عصبی ساختارمند
۳۴	TRADES ۵-۳-۲
۳۵	۱۶-۳-۲ روش های اثبات پذیر مقاوم
۳۵	۱۷-۴-۲ تاثیر یادگیری مقاوم
۳۶	۱۸-۵-۲ روش های حذف اثر دسته ای
۳۷	۱۹-۵-۲ یادگیری مقابله ای عمیق در معیارهای زیستی
۳۷	۲۰-۵-۲ شبکه های باقیمانده ای یکسان کننده توزیع
۳۸	۲۱-۵-۲ حذف اثر دسته ای به کمک رمزگذاری مستقل از دسته
۳۹		۲۲-۳-۲ روش پیشنهادی

۴۳	۴ ارزیابی و نتایج
۴۳	۱-۱ مجموعه داده‌ها
۴۵	۱-۱-۱ داده‌های مسئله
۴۶	۲-۱ معیارهای ارزیابی
۴۶	۱-۲-۱ ارزیابی روش درهم
۴۶	۱-۲-۲ ارزیابی روش کنارگذاشتن تکی
۴۷	۳-۱ جزئیات پیاده سازی
۴۹	۴-۱ ارزیابی دقت
۵۱	۵-۱ ارزیابی کیفی
۵۲	۶-۱ تاثیر بودجه انحراف بر بیش‌بازش مدل‌ها
۵۵	۵ نتیجه‌گیری
۵۶	۱-۱ کارهای پیش رو
۵۶	۱-۱-۱ استفاده از حمله‌های متفاوت
۵۷	۱-۱-۲ یکسان سازی دسته‌ها با تعداد گام‌های متغیر

فهرست شکل‌ها

۱-۱ نمونه اصلی از مجموعه داده و نویز هدف‌دار	۱۲
۱-۲ انحراف تصویر با دو اندازه مختلف	۱۲
۱-۳ انحراف تابلو راهنمایی و رانندگی	۱۳
۱-۴ عینک فریب‌دهنده دستگاه‌های تشخیص دهنده	۱۳
۱-۵ توزیع داده‌ها در راستای دو مولفه اصلی	۱۵
۱-۶ نمونه‌ای از اثر دسته‌ای ایجادشده در تصویر OPG دندانپزشکی	۱۶
۲-۱ بررسی عامل بیش‌برازش در بوجود‌آمدن نمونه‌های تخاصمی	۲۰
۲-۲ بررسی خطی بودن شبکه‌های عمیق	۲۰
۲-۳ تغییرات ورودی در یک راستای تصادفی و بررسی اشکال خطی بودن	۲۱
۲-۴ مقایسه توابع نتیجه‌گیری در حالت مطلوب و نامطلوب	۲۲
۲-۵ مقایسه حالت‌های مختلف برای آموزش مدل و بررسی مقاوم بودن مدل	۲۳
۲-۶ تغییرات در راستای روش گرادیان سریع و راستای تصادفی	۲۴
۲-۷ مراحل الگوریتم نزول گرادیان افکنده	۲۶
۲-۸ مراحل الگوریتم حمله مرز تصمیم	۲۹
۲-۹ تعداد کوئری مورد نیاز برای حالت هدفمند و غیرهدفمند	۳۰
۲-۱۰ نمونه‌ای از حمله همگانی انجام شده	۳۱

۲-۱۱ رابطه استفاده شده برای آموزش شبکه معلم	۳۳
۲-۱۲ نحوه اجرای الگوریتم تقطیر دفاعی	۳۳
۲-۱۳ نحوه اجرای الگوریتم تقطیر دفاعی	۳۴
۲-۱۴ ویژگی‌های مورد توجه مدل‌های عادی و مقاوم	۳۶
۳-۱ الگوریتم حذف اثر دسته‌ای	۴۱
۳-۲ تاثیر روش پیشنهادی	۴۱
۴-۱ دقیق مدل‌های استاندارد نسبت به حمله‌های تخاصمی	۴۹
۴-۲ توزیع داده‌ها پس از اجرای حمله	۵۲
۴-۳ توزیع داده‌ها قبل از اجرای حمله	۵۲

فهرست جدول‌ها

۴۴	۱-۱ ویژگی‌های مجموعه داده
۴۸	۲-۲ معماری شبکه پیچشی استفاده شده
۵۰	۳-۳ دقت مدل‌ها، مجموعه داده SA
۵۰	۴-۴ دقت مدل‌ها، مجموعه داده DEAP
۵۴	۵-۵ بیش‌بازش مدل‌ها، مجموعه داده SA
۵۴	۶-۶ بیش‌بازش مدل‌ها، مجموعه داده DEAP

فصل ۱

مقدمه

۱-۱ مفاهیم اولیه

۱-۱-۱ نمونه‌های تخاصمی^۱

شبکه‌های عمیق^۲ دسته‌ای از مدل‌های یادگیری ماشین هستند. هدف از روش‌های یادگیری عمیق، یادگیری سلسله مراتبی ویژگی‌ها با استفاده از استخراج ویژگی‌های سطح بالاتر است که از ترکیب ویژگی‌های سطح پایین تر تشکیل شده است. یادگیری خودکار، ویژگی‌ها در سطوح متعدد انتزاعی استخراج می‌کند و سپس این ویژگی‌هارا ترکیب می‌کند و به سیستم اجازه می‌دهد تا عملکردهای پیچیده‌ای را که نقشه و توزیع ورودی را به خروجی ترسیم می‌کنند مستقیماً از داده‌ها بیاموزد ، بدون اینکه به ویژگی‌های ساخته شده دست انسان نیاز داشته باشد [۱].

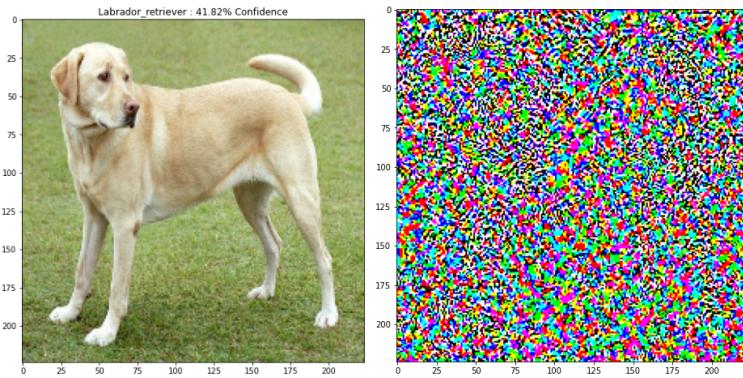
این موضوع که به ویژگی‌های دست انسان نیاز ندارد مزايا و معایبی خواهد داشت که در ادامه به آن خواهیم پرداخت.

به نمونه‌های خصمانه برای فریب‌دادن شبکه‌های عمیق نمونه‌های تخاصمی گفته می‌شود، این نمونه‌ها از جنس مجموعه داده می‌باشد. این نمونه‌ها با استفاده از نمونه‌های مجموعه داده و اضافه کردن نویزی خاص بدست می‌آیند. این نویزها معمولاً به چشم انسان ناچیز به نظر می‌رسند ولی می‌توانند با دقت خوبی تمام شبکه‌های یادگیری عمیق را فریب‌دهند. برای بدست آوردن یک نمونه تخاصمی روش‌های

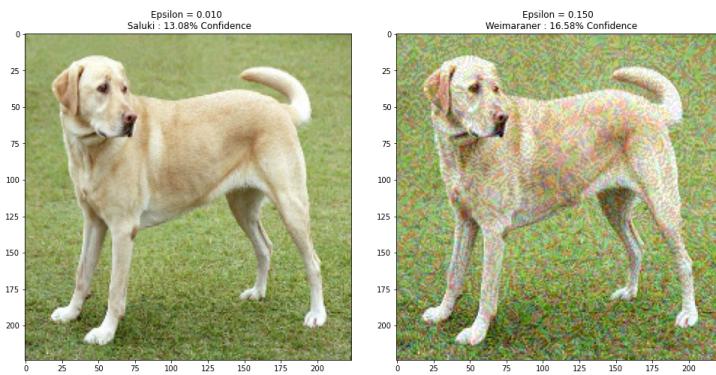
¹ Adversarial Examples

² Deep Neural Network

مختلفی وجود دارد که به حمله تخاصمی معروف هستند. تمامی آن‌ها بر این تکنیک افزایش هزینه مدل استوار هستند [۲].



شکل ۱-۱: نمونه اصلی از مجموعه داده و نویز هدف‌دار [۳]



شکل ۱-۲: انحراف تصویر با دو اندازه مختلف [۳]

همان‌طور که در اشکال ۱-۱ و ۱-۲ مشاهده می‌کنید اگرچه اندازه نویز کم می‌باشد و به چشم انسان قابل تشخیص نیست اما مقدار قابل توجهی خروجی شبکه را تغییر می‌دهد و باعث فریب آن می‌شود. در نگاه اول ممکن است نمونه‌های تخاصمی دور از ذهن و کم اهمیت به نظر برسند، اما زمانی که کاربرد وسیع مدل‌های عمیق را در زندگی روزمره و آینده نزدیک در نظر بگیریم، خطرات آن‌ها واضح‌تر می‌شود:

۱. می‌توان با استفاده از برچسب‌هایی که بر روی جاده قرار داده شده‌اند ماشین‌های خودران را فریب‌داد تا در مسیر مخالف حرکت کنند [۴].



شکل ۱-۳: انحراف تابلو راهنمایی و رانندگی [۴]

۲. عینک‌هایی که با چاپگر سه‌بعدی تولید شده‌اند و سیستم‌های تشخیص چهره را فریب می‌دهند [۵].



شکل ۱-۴: عینک فریب‌دهنده دستگاه‌های تشخیص چهره [۵]

۳. در روش‌های یادگیری تقویتی^۳ استفاده از نمونه‌های تخاصمی در ورودی و سیاست^۴ می‌تواند باعث کاهش محسوس عملکرد عامل شود [۶].

۴. فرمان‌های تخاصمی برای دستیاران صوتی که برای انسان غیرقابل تشخیص است [۷].

³Reinforcement Learning

⁴Policy

در تمامی موارد بالا می‌توان مشاهده نمود که نمونه‌های تخاصمی چقدر می‌توانند خطرناک باشند، از این رو می‌بایست شبکه‌های عمیق را قدرتمندتر نمود که در مقابل این نمونه‌های تخاصمی مقاوم باشند. در تمامی موارد بالا برای تولید نمونه تخاصمی نیاز به معماری شبکه و وزن‌های مدل یادگرفته شده داریم پس به نظر می‌رسد با پنهان نگه داشتن مدل و وزن‌های آن می‌توان این شبکه‌هارا مقاوم کرد در حالی که امروزه پژوهش‌های انجام شده حاکی از آن است که می‌توان حملات جعبه‌سیاه^۵ به این مدل‌ها انجام داد به صورتی که تنها با داشتن ورودی و خروجی شبکه در طی چند مرحله می‌توان مدل خصمانه تولید نمود [۸].

علاوه بر مفاهیم ارائه شده نتایج نشان می‌دهد که نمونه‌های خصمانه برای یک مدل اغلب نمونه خصمانه برای مدل‌های دیگر است که به این خاصیت انتقال پذیری^۶ گفته می‌شود. حال با توجه به این گزاره می‌توان نتیجه گیری نمود که اگر دسترسی به مدل ممکن نباشد می‌توان نمونه خصمانه بر روی مدل مشابه دیگری شبیه‌سازی نمود و از آن در انحراف مدل اصلی استفاده کرد [۱].

۲-۱-۱ یادگیری تخاصمی^۷

یادگیری ماشین تخاصمی^۸ یک روش یادگیری ماشین است که هدف آن فریب مدل‌های یادگیری ماشین با ارائه ورودی‌های فریب‌دهنده است، در نتیجه هم شامل تولید و هم شناسایی نمونه‌های تخاصمی می‌شود – ورودی‌هایی که به خصوص برای فریبدادن مدل مسئله‌های دسته بندی^۹ ایجاد شده‌اند. انواع حمله در یادگیری تخاصمی:

۱. حمله‌های جعبه‌سفید^{۱۰}: حملاتی هستند که در آن مهاجم به مدل هدف (معماری و وزن‌های مدل) دسترسی کامل دارد [۹].

۲. حمله‌های جعبه‌سیاه^{۱۱}: حملاتی هستند که در آن مهاجم به مدل دسترسی ندارد و تنها می‌تواند به ازای ورودی مطلوب خروجی را مشاهده نماید [۹].

⁵Black-Box Attack

⁶Transferability

⁷Adversarial Learning

⁸Adversarial Machine Learning

⁹Classification

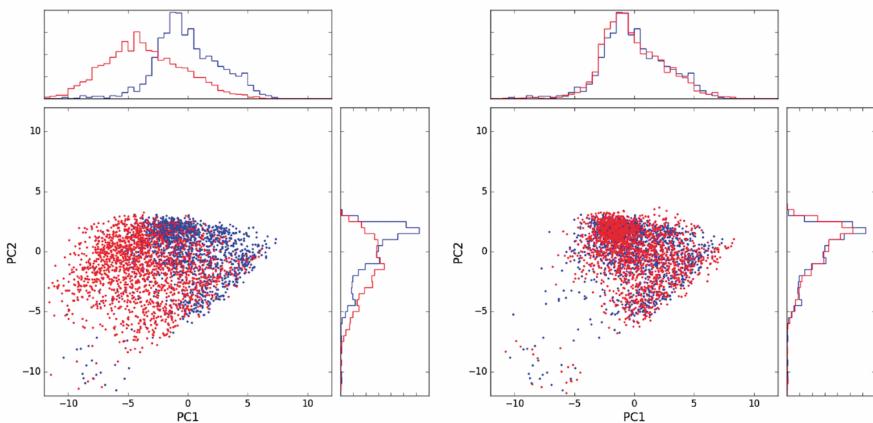
¹⁰White-Box

¹¹Black-Box

۱-۱-۳ اهمیت موضوع

مدل‌های یادگیری ماشینی به خودی خود در برابر نمونه‌های تخاصمی مقاوم نیستند. در مدل‌های یادگیری ماشین فرض بر این است که توزیع داده‌ها در زمان آموزش و تست یکسان است و دقیقی که در زمان آموزش بدست می‌آید می‌تواند تخمین خوبی برای دنیای واقعی و زمان تست باشد.

نمونه‌های تخاصمی توزیع متفاوتی نسبت به داده‌ها دارند، در نتیجه باعث نقض فرض بالا می‌شوند. از آنجایی که هدف کمینه کردن خطای داده‌های آموزش بوده است نمی‌توان هیچ اطمینانی برای کمبودن خطای داده‌های آزمون داشت. روش‌های مختلفی برای مقاوم کردن مدل‌های یادگیری ماشین ارائه شده است که معمولاً از یک یا چند حمله در طی فرایند آموزش استفاده می‌کنند و سعی می‌کنند تا حد خوبی مدل آموزش دیده را در برابر حملات مقاوم کنند [۱۰].



شکل ۱-۵: توزیع داده‌ها در راستای دو مولفه اصلی اول برای چپ: داده‌ها با اثر دسته‌ای، راست: بعد از حذف اثر دسته‌ای. نقاط آبی و قرمز مربوط به دو دسته مختلف می‌باشند. دیده می‌شود که بعد از حذف اثر دسته‌ای، توزیع دسته‌ها شباهت بیشتری به یکدیگر پیدا کرده اند [۱۰].

۱-۱-۴ اثر دسته‌ای^{۱۲}

داده‌های زیستی در هنگام جمع آوری می‌توانند تحت تاثیر عوامل متعددی قرار بگیرند که باعث شود خطاهای بسیاری در آن‌ها هرچند کوچک رخ دهد. در مجموعه داده‌های زیستی دو منبع برای تولید اثر دسته‌ای داریم: تنوع بیولوژیکی (ناشی از تفاوت در جمعیت‌ما) – تنوع غیربیولوژیکی (ناشی از تفاوت در دستگاه‌های اندازه گیری).

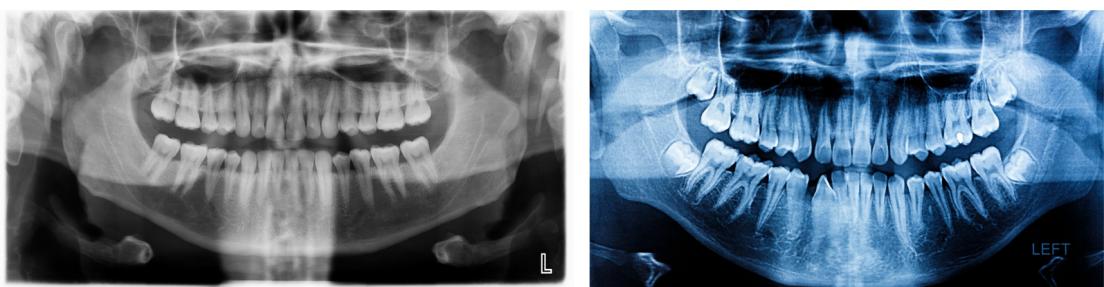
¹²Batch Effect

در اینجا چند مشکل وجود دارد. اولین مشکل مربوط به یادگیری میانبر^{۱۳} است. فرض کنید ما مجموعه داده‌ای از تصاویر سی‌تی قفسه سینه از بیمارستان A و بیمارستان B برای تشخیص سرطان ریه داریم. بیمارستان A در یک شهر آلوده است و احتمال ابتلا به سرطان ریه بسیار بیشتر از شهر B است. دستگاه سی‌تی در بیمارستان A نیز تمايل دارد تصاویر روشن‌تری نسبت به بیمارستان B تولید کند. اکنون زمانی که ما یک شبکه عصبی پیچشی^{۱۴} را بر روی مجموعه داده آموخته می‌دهیم، برای پیش‌بینی سرطان ریه، مدل یاد می‌گیرد که تصاویر روشن‌تر را با سرطان مرتبط کند. مطمئناً اگر روی همین مجموعه داده اعتبار سنجی کنیم ممکن است دقت پیش‌بینی مناسبی داشته باشیم. اما وقتی مدل خود را در جای دیگری مستقر کنیم ممکن است دقت مدل به شدت افت کند.

چرا داده‌های خود را غنی‌تر نمی‌سازیم؟^{۱۵}

در این پژوهش سعی بر این است که داده‌های مسئله را به نوعی افزایش دهیم به طوری که مدل را از یادگیری ویژگی‌های غیرمطلوب منصرف کنیم، این ویژگی‌ها عمومی در سطح دسته‌ای شدن مجموعه داده هستند و با داده افزایی‌های کلاسیک عملاً اثر این ویژگی‌ها از بین نمی‌رود. به طور مثال این ویژگی‌ها می‌توانند نویز، رنگ پس زمینه و ... باشند.

گاهی از اوقات زیرتوزیع‌های مختلف را داریم ولی عامل تمایز این توزیع‌ها در دسترس و توضیح پذیر نمی‌باشد در نتیجه نمی‌توانیم به راحتی مجموعه داده‌هارا غنی سازیم.



شکل ۱-۶: نمونه‌ای از اثر دسته‌ای ایجاد شده در تصویر OPG دندانپزشکی [۱۱]

¹³Shortcut Learning

¹⁴Convolutional Neural Network

¹⁵Data Augmentation

۱-۲ طرح مسئله

اثر دسته‌ای به عنوان یک نویز در داده‌های زیستی، باعث می‌شود که توزیع داده در دسته‌های مختلف، توزیع‌های متفاوتی داشته باشند و مدل‌های یادگیری عمیق نتوانند به صورت بهینه بر روی آن‌ها عمل کنند و یادگیری روی آن‌هارا با مشکل مواجه سازد. همچنین می‌توان فرض کرد که اندازه این نویزها معمولاً کوچک است و رفتار آن‌ها نیز نزدیک به تابع همانی است. این رفتار را می‌توان بسیار شبیه به نمونه‌های تخاصمی دید که با تغییرات کوچکی در ورودی، باعث به اشتباه افتادن مدل می‌شوند. یادگیری تخاصمی به عنوان پاسخی برای آسیب تخاصمی^{۱۶} سعی می‌کند مدلی را یاد بگیرد که نسبت به حملات تخاصمی مقاوم باشد. هدف این پایان‌نامه، استفاده از چهارچوب یادگیری تخاصمی به عنوان روشی موثر برای مدل‌سازی و حذف اثر دسته‌ای است. نشان خواهیم داد که روش ارائه شده، قادر است دقت مدل‌های یادگیری ماشین را در حضور اثرات دسته‌ای بهبود بخشد و توزیع دسته‌های مختلف را یکسان کند.

^{۱۶} Adversarial Vulnerability

۱-۳ مسئله اسب باهوش

این مسئله یک مسئله معروف در تئوری یادگیری ماشین است.

در این مسئله صاحب یک اسب متوجه می‌شود اسبش توانایی محاسبه عملیات ریاضی را دارد. بدین صورت که پس از هر پرسش از اسب، اسب با کوبیدن سم بر روی زمین حاصل جواب را اعلام می‌کند مثلاً اگر صاحب هست سوال $3 + 2$ را از اسب پرسیده باشد اسب ۵ بار سم بر زمین می‌کوبد.

این آزمایش چندین بار در مکان‌های عمومی اجرا می‌شود و اسب در هر پرسش پاسخ صحیح را به مسئله می‌دهد. پس از آن صاحب اسب برای کشف چگونگی حل مسئله توسط این اسب آن را به مکانی خلوت می‌برد و بر روی سر خودش کیسه‌ای می‌کشد که اسب قیافه او را نمی‌بیند و متوجه تغییرات چهره او نشود، سپس از اسب سوال می‌پرسد و مشاهده می‌کند اسب دچار اشتباه شده و به صورت مداوم در حال کوبیدن سم است!

از این آزمایش معروف می‌توان این نتیجه را گرفت که اسب با توجه به حالت چهره افراد در جمعیت جواب مسئله را پیدا می‌کرده است و واقعاً توانایی حل مسئله را ندارد، چنان مفهومی در مدل‌های یادگیری ماشین می‌تواند وجود داشته باشد یعنی مدل به جای یادگیری ویژگی‌های مد نظر انسان و منطقی به دنبال ویژگی‌های دیگری برود که خواسته مسئله به بهترین نحو برآورده شود [۱۲].

یکی از راه‌های مقابله با این موضوع استفاده از توزیع‌های مختلف یک مسئله در زمان آموزش و تست است.

فصل ۲

بررسی کارهای پیشین

در این فصل ابتدا به دلایل ایجاد آسیب‌پذیری در مدل‌های یادگیری ماشین می‌پردازیم، پس از آن به حملات معروف در این زمینه اشاره خواهیم کرد.

۱-۲ عوامل آسیب‌پذیری مدل‌های یادگیری ماشین

در این قسمت به مشکلاتی که باعث بروز آمدن آسیب‌پذیری در این مدل‌ها می‌شوند می‌پردازیم.

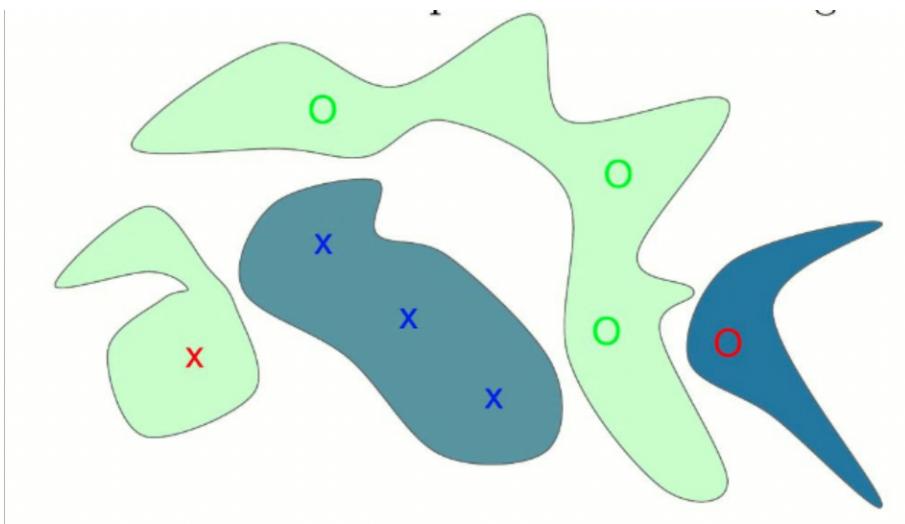
۱-۱-۲ بیش‌برازش

اولین فرضیه مطرح شده این است که در شبکه‌های عمیق^۱ احتمال بیش‌برازش^۲ داده‌ها زیاد است در نتیجه این مشکل می‌تواند باعث بروز آسیب‌پذیری شود. در شکل بالا ۳ نمونه‌ی دایره وسط و ۳ نمونه ضربدر وسط، نمونه‌های آموزش ما بوده اند و شبکه عمیق بیش‌برازش شده مرزهای رنگی مربوطه را برای هر کلاس تعیین کرده است، حال به وضوح می‌توان با تغییر در برخی نمونه‌های آموزش نمونه‌های تخاصمی ایجاد کرد.

از شکل و بررسی‌های بالا می‌توان نتیجه‌گیری کرد که عامل ایجاد نمونه‌های تخاصمی، بیش‌برازش مدل‌ها است اما این نمونه‌های تخاصمی این ویژگی را دارند که در بقیه مدل‌های یادگیری ماشین نیز می‌توانند

¹Deep Neural Networks

²Overfitting

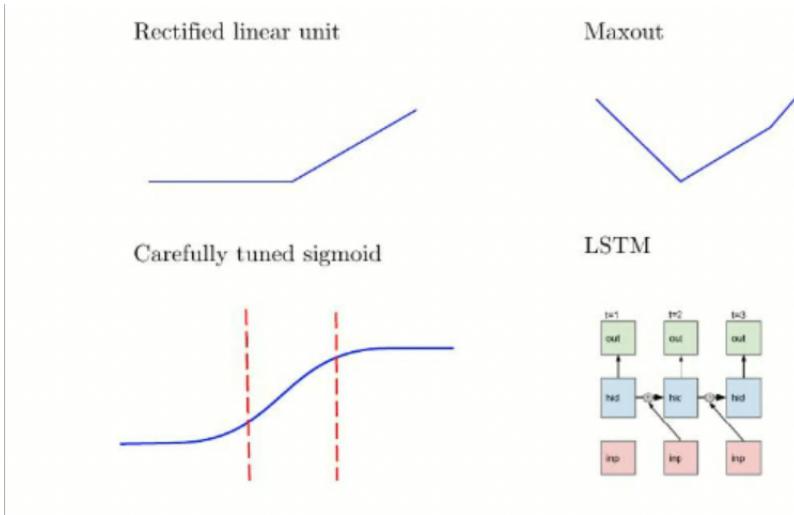


شکل ۲-۱: بررسی عامل بیشبرازش در بوجود آمدن نمونه‌های تخاصمی [۱۳]

به عنوان نمونه تخاصمی عمل کنند و مدل را فریب دهنند در نتیجه این فرض رد می‌شود [۱۳].

۲-۱-۲ خطی بودن مدل‌های یادگیری ماشین

اکثر مدل‌های شبکه‌های عمیق از واحد های خطی در بدنه خود استفاده می‌کنند.



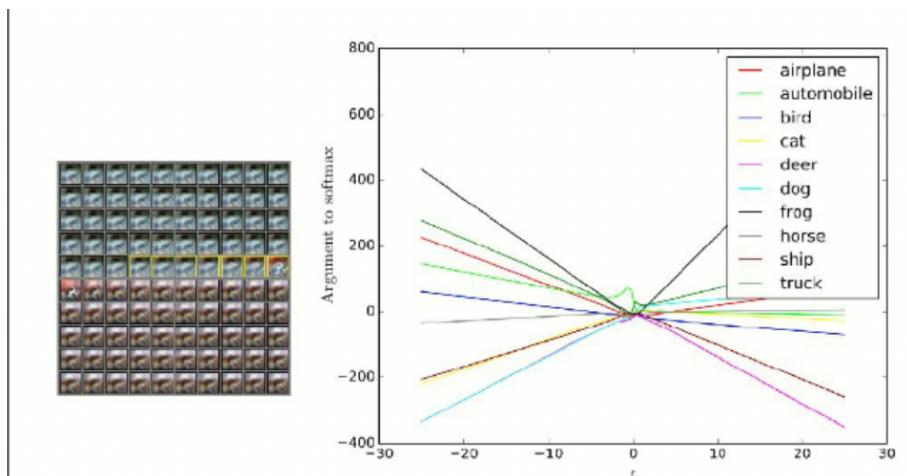
شکل ۲-۲: بررسی خطی بودن شبکه‌های عمیق [۱۳]

همان‌طور که در شکل بالا مشاهده می‌کنید واحد RELU عملکرد خطی تکمای دارد^۳ و تابع-Sig-

³Piecewise Linear

نیز چون اکثرن ما در نواحی غیراشعاع هستیم به صورت تقریباً خطی عمل می‌کند، هم چنین در شبکه‌های LSTM^۴ نیز بخش‌های مخفی^۵ و خروجی تابع از عملگرهای خطی و توابع خطی استفاده می‌کنند [۱۳].

از بررسی‌های بالا می‌توان این نتیجه‌گیری را نمود که مدل‌های یادگیری ماشین اکثرن به صورت تکه‌ای خطی هستند.



شکل ۲-۳: تغییرات ورودی در یک راستای تصادفی و بررسی اشکال خطی بودن [۱۳]

شکل بالا یک نمونه از مجموعه داده CIFAR می‌باشد. همان‌طور که در شکل بالا مشاهده می‌کنید توزیع داده‌ها حول نقطه $\epsilon = 0$ است و شبکه در حول این نقطه به درستی خروجی را مقدار اتومبیل قرار داده است اما به فاصله گرفتن از حول این نقطه و عملکرد خطی مدل‌های یادگیری ماشین می‌توان مشاهده کرد که در نقاطی از فضای نمایانگر مقداری در آن نقطه نبوده است شبکه تصمیم‌قاطعی درباره خروجی در این نقطه دارد که این همان آسیب‌پذیری مورد بحث است [۱۳].

۲-۱-۳ تصمیم‌گیری در نقاط نامعلوم

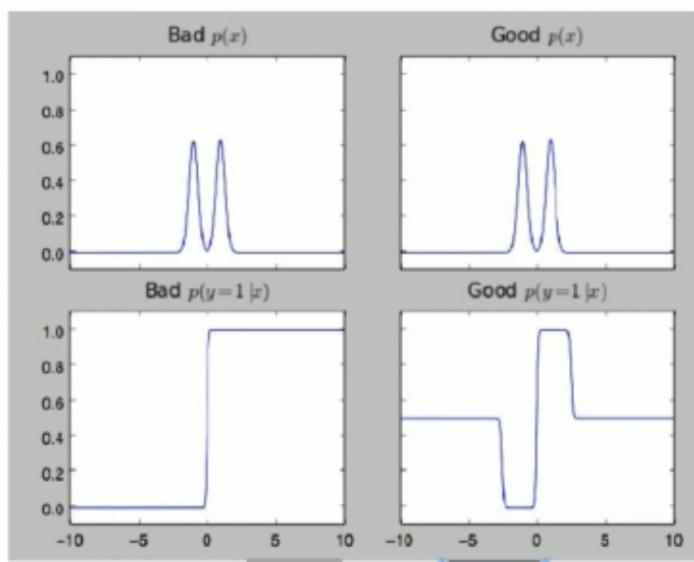
در بررسی دلایل آسیب‌پذیری شبکه‌های عمیق این مورد اصلی ترین می‌باشد، شبکه‌های عمیق معمولاً در خمیده^۶ توزیع داده‌های آموزش نتایج منطقی و مطلوب دارند و هرچه از این ناحیه دورتر می‌شویم تصمیمات با درجه قطعیت بیشتری گرفته می‌شوند که این مشکل اساسی است زیرا درباره نواحی ناشناخته

⁴Long Short-Term Memory

⁵Hidden

⁶Manifold

خروجی شبکه می‌بایست عدم قطعیت را نشان دهد [۱۲]. در نگاه احتمالاتی به قضیه $p(y|x)$ مهم است،



شکل ۲-۴: مقایسه توابع نتیجه‌گیری در حالت مطلوب و نامطلوب [۱۳]

حال همان‌طور که در شکل بالا مشاهده می‌کنید انتخاب (x) مناسب می‌تواند به شکل موثری جلوی ایجاد نمونه تخاصمی را بگیرد.

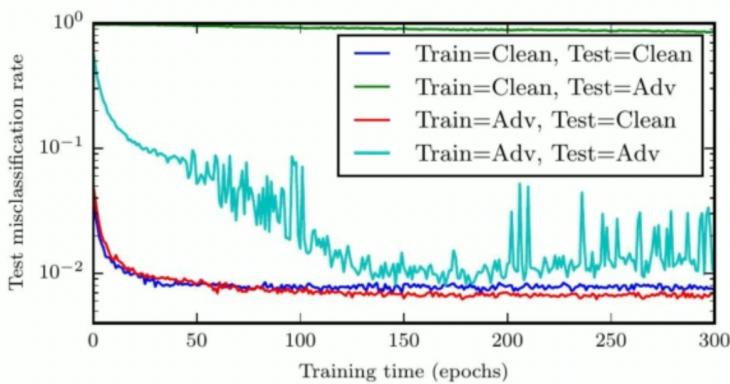
مثلا در شکل بالا اگر (x) یک توزیع از ترکیب دو توزیع نرمال^۷ باشد خروجی در نواحی دورتر از صفر تصمیم‌گیری با قاطعیت دارد ولی اگر (x) از توزیع لابلسین^۸ انتخاب شود، در ناحیه نزدیک به صفر پیش‌بینی‌های درستی داریم و ناحیه‌های دورتر را عدم قطعیت می‌دانیم که این مطلوب مسئله است [۱۳].

⁷Normal Distribution

⁸Laplace Distribution

۴-۱-۲ تفاوت توزیع ها در زمان آموزش و تست

برای بررسی مقاوم بودن مدلی که آموزش داده می شود از نمونه های تخاصمی استفاده می شود.



شکل ۲-۵: مقایسه حالت های مختلف برای آموزش مدل و بررسی مقاوم بودن مدل [۱۴]

در شکل بالا زمانی که داده های آموزشی از نمونه های تخاصمی نباشند و توزیع داده های تست از نمونه های تخاصمی آمده باشد نتیجه اصلا خوب نیست و درصد اشتباه در دسته بندی بسیار بالاست. حال اگر علاوه بر نمونه های واقعی، نمونه های تخاصمی را به آموزش مدل اضافه کنیم می توان مشاهده کرد که مدل مقاوم تری خواهیم داشت و اشتباه در دسته بندی به شدت کاهش پیدا می کند [۱۵]. اما مقایسه اصلی در واقع بین نمودار سبز و آبی کمرنگ است که ما مدل را بر روی توزیع داده های تخاصمی تست می کنیم. در نتیجه زمانی که نمونه های تخاصمی در مجموعه آموزشی ما هستند نتایج مطلوب تر است.

۲-۲ انواع حمله های تخاصمی

از آنجایی که روش های آموزش تخاصمی، از نمونه های تخاصمی در فرایند آموزش استفاده می کنند، در ابتدا انواع مختلف حمله های تخاصمی که روش تولید نمونه های تخاصمی هستند را بررسی می کنیم. سپس الگوریتم های مهم یادگیری تخاصمی را بررسی می کنیم. در انتها به روش هایی که تاکنون برای حل مشکل اثر دسته ای مطرح شده اند، می پردازیم.

در این روش ها، هدف این است که به ازای ورودی داده شده x که به کلاس x تعلق دارد، مقدار نویز $\Delta \in \delta$ به گونه ای پیدا شود که اندازه δ تا حد امکان کوچک باشد و $\delta + x$ توسط مدل به کلاسی جز y

دسته‌بندی شود. Δ مجموعه تمام نویز‌های معتبر را نشان می‌دهد که معمولاً با شرایطی مانند داشتن ℓ_p محدود مشخص می‌شوند [۱۶].

$$x_{adv} = x + \arg \min_{\delta} \{ \|\delta\|_p : f(x) \neq y \} \quad (1-2)$$

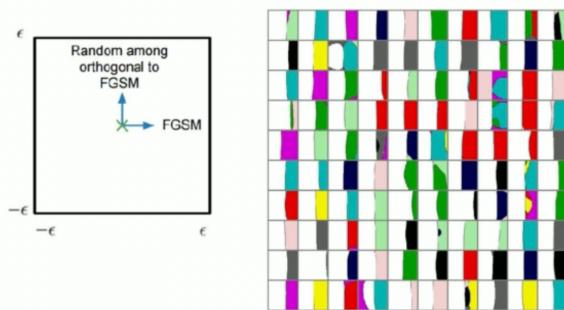
در رابطه بالا $f(x)$ خروجی مدل به ازای ورودی x را نشان می‌دهد.

۲-۱ روشن علامت گرادیان سریع

در این روش سعی می‌شود با گرفتن گرادیان تابع هزینه نسبت به ورودی راستایی را پیدا کرد که با حرکت در آن راستا بیشترین افزایش را در تابع هزینه داشته باشیم.

$$x_{adv} = x + \epsilon * sign(\nabla_x L(x, y)) \quad (2-2)$$

در این روش اندازه‌ای که نویز تخاصمی در آن محدود می‌شود، اندازه بی نهایت (ℓ_∞) است. همان‌طور که در رابطه بالا مشاهده می‌شود مقدار نویز تخاصمی به ϵ محدود شده است، این روش با وجود اینکه بسیار ساده است و محاسبه آن سریع است، دقت خوبی هم در تولید نمونه‌های تخاصمی دارد. اما برای آموزش مدل‌ها مناسب نیست. زیرا نمونه تولید شده توسط این حمله پس از چند مرحله آموزش، توسط مدل یاد گرفته می‌شوند و در واقع مدل به نوعی بر روی این حمله دچار بیش‌برازش می‌شود.



شکل ۲-۶: بررسی تغییرات خروجی مدل بر حسب تغییرات در راستای گرادیان روش علامت گرادیان سریع و راستای تصادفی [۱۷]

همان‌طور که در شکل بالا قابل مشاهده است تغییر در ورودی همیشه منجر به تولید نمونه تخاصمی نمی‌شود و فقط حرکت در برخی راستاها باعث تولید نمونه تخاصمی می‌شود. در شکل اگر در راستای

روش علامت گرادیان سریع حرکت کنیم مقدار خروجی تغییر می‌کند (به سمت راست حرکت کنیم) ولی اگر به سمتی تصادفی حرکت کنیم مقدار خروجی تغییری نخواهد کرد [۱۷].

L-BFGS ۲-۲-۲

این روش سعی دارد رابطه (۲-۱) را به یک مسئله بهینه‌سازی تبدیل کند و آنرا با استفاده از الگوریتم BFGS حل نماید، در این الگوریتم سعی می‌شود علاوه بر کوچک نگه داشتن ℓ_2 نویز، نمونه با اضافه شدن نویز به کلاسی دیگری تعلق گیرد.

$$\|\delta\|_2 + Loss(x + \delta, y') \quad (3-2)$$

این روش دقت بسیار بالایی دارد اما محاسبات آن زمانبر است و نیازمند منابع حافظه زیادی است.
دقت روش‌های تخاصمی با استفاده از رابطه زیر سنجیده می‌شود.

$$SuccessRate = \frac{\text{Successful Adversarial Examples}}{\text{All Adversarial Examples}} \quad (4-2)$$

نمونه‌های موفق نمونه‌هایی هستند که مدل را فربین داده باشند [۱۸].

۳-۲-۲ حمله یک گامی هدف دار

در حملات غیرهدف دار مانند حمله علامت گرادیان سریع، فقط می‌خواهیم نمونه به کلاس دیگری تعلق پیدا کند و تنها چیزی که مهم است این است که مدل این نمونه را درست دسته‌بندی نکند، اما در حمله هدف دار می‌خواهیم تابع هزینه را به ازای این ورودی کاهش دهیم تا مدل این ورودی را به کلاسی که مدل نظر ماست احتمال بالایی دهد [۱۹].
رابطه مورد استفاده از این حمله در زیر آمده است.

$$x_{adv} = x - \epsilon * sign(\nabla_x L(x, y)) \quad (5-2)$$

۴-۲-۲ روشن تکرارشونده ساده

این روش یک گسترش^۹ از روش علامت گرادیان سریع^{۱۰} است، در این روش چند بار از رابطه (۲-۲) استفاده می‌کنیم و در هر مرحله از خروجی مرحله قبلی نمونه تخاصمی جدید را می‌سازیم.

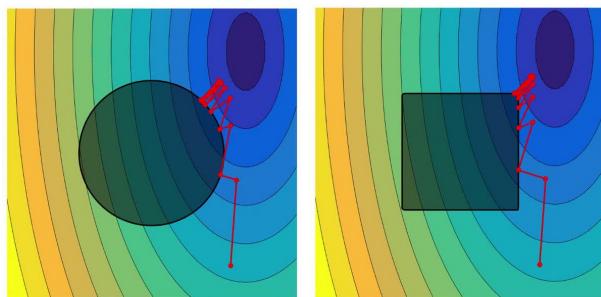
$$x_{\cdot} = x, x_{adv,t} = x_{adv,t-1} + \alpha * sign(\nabla_x L(x_{adv,t-1}, y)) \quad (6-2)$$

در رابطه بالا α اندازه گام در هر مرحله است که معمولاً از ϵ کوچکتر است. این روش با انتخاب گام‌های کوچک تر و تعداد گام‌های بیشتر می‌تواند نمونه‌های بهتری را پیدا کند. این روش نیازمند این است که ورودی چندین بار به مدل داده شود و گرادیان‌ها محاسبه شوند و در واقع هزینه آن به تعداد تکرارهاییش بستگی دارد. این روش برای تکرارهای زیاد نتایج بسیار خوبی تولید می‌کند [۱۹].

۴-۲-۳ نزول گرادیان افکنده شده

تفاوت این روش با روش قبلی در این است که در هر مرحله، نمونه تخاصمی بدست آمده توسط یک معیار فاصله^{۱۱} به یک همسایگی ثابت حول x افکنده می‌شود. از معیارهای فاصله پرکاربرد در این روش می‌توان به معیارهای ℓ_2 و ℓ_{∞} اشاره کرد.

این روش به عنوان پایه بسیاری از حمله‌های تخاصمی جدیدتر مورد استفاده قرار گرفته است و تاکنون تمام دفعه‌های موفق در این زمینه، به نوعی از این حمله برای ساختن نمونه‌های تخاصمی خود استفاده کرده‌اند [۲۰].



شکل ۲-۷: مراحل الگوریتم نزول گرادیان افکنده [۲۱]

⁹Extension

¹⁰FGSM

¹¹Distance Measure

۱. شروع از یک انحراف تصادفی در فاصله کمتر از ℓ_p تا نمونه واقعی
۲. محاسبه گرادیان در بهترین راستا برای افزایش تابع هزینه و حرکت در آن راستا
۳. اگر فاصله نتیجه بدست آمده تا نمونه واقعی از ℓ_p بیشتر باشد می‌بایست نتیجه بدست آمده را بر روی نزدیکترین نقطه فاصله عمود کنیم.
۴. گام‌های ۲ و ۳ را تا مرز همگرایی ادامه می‌دهیم.

۶-۲-۲ نقشه برجستگی مبتنی بر ماتریس ژاکوبین

در این روش سعی می‌شود با استفاده از ماتریس ژاکوبین پیشینی مدل نسبت به ورودی، مولفه‌هایی از ورودی پیدا شوند که تغییر در آن‌ها بیشترین تغییر در خروجی را ایجاد می‌کند. سپس کمترین تعداد مولفه‌هایی که بیشترین تاثیر در خروجی را دارند انتخاب می‌شوند و مقدار آن‌ها را با توجه به مثبت و منفی بودن و تاثیر حداکثری کم و زیاد می‌کنند. این روش سعی در کم نگه داشتن ℓ دارد. یعنی کمترین درایه ممکن را تغییر دهد تا خروجی مدل تغییر کند. این روش پایه‌ی دسته‌ای از روش‌های است که سعی دارند معیار اندازه ℓ نویز را کمینه کنند. این روش‌ها با توجه به هزینه‌ی بالای محاسبه و مطلوب نبودن معیار اندازه در برخی شرایط، کمتر مورد توجه قرار گرفته‌اند [۲۲].

مدل‌های جعبه سفید به دو دسته‌ی هدف مند و غیرهدف مند تقسیم می‌شوند. مدل‌های جعبه سیاه به دو دسته‌ی مبتنی بر امتیاز و مبتنی بر مرز تقسیم می‌شوند. مدل‌های مبتنی بر مرز تنها هدف تغییر عضو بیشینه در انتساب دسته هارا هدف قرار می‌دهند اما مدل‌های مبتنی بر امتیاز سعی در تغییر امتیازات همه دسته‌ها دارند [۲۳].

Deep-Fool ۷-۲-۲

این الگوریتم، یک الگوریتم جعبه سفید غیرهدفمند است یعنی فقط قصد فریب شبکه‌های عمیق را دارد. روش این الگوریتم به این صورت است که ابتدا مرزهای دسته‌بندی را بین کلاس‌های مختلف پیدا می‌کند و سپس نمونه ورودی را به یکی از این مرزها نگاشت کرده و سپس با حرکت بسیار کوچکی در طرف دیگر مرز باعث به اشتباہ انداختن شبکه‌های عمیق می‌شود.

در بسیاری از مواقع مرز تصمیم‌گیری غیرخطی می‌باشد که در این صورت این مدل مرزهای خطی فرض می‌کند و در صورت رد نشدن از مرز دوباره این عمل تخمین مرز را انجام می‌دهد. این الگوریتم در مقایسه با FGSM بسیار کارتر می‌باشد و نویز ایجاد شده به چشم انسان کم اثر تر است [۲۴].

حمله‌های جعبه سیاه ۸-۲-۲

همان‌طور که اشاره شد دسته دیگری از حملات هستند که نیازی به دانستن ساختار داخلی مدل ندارند و صرفا با دادن ورودی به مدل و دیدن خروجی آن کار می‌کنند. این روش‌ها معمولاً مدلی مشابه مدل مورد نظر می‌سازند و با استفاده از روش‌های قبلی نمونه‌های تخاصمی را تولید می‌کنند [۲۵].

حمله مرز تصمیم ۹-۲-۲

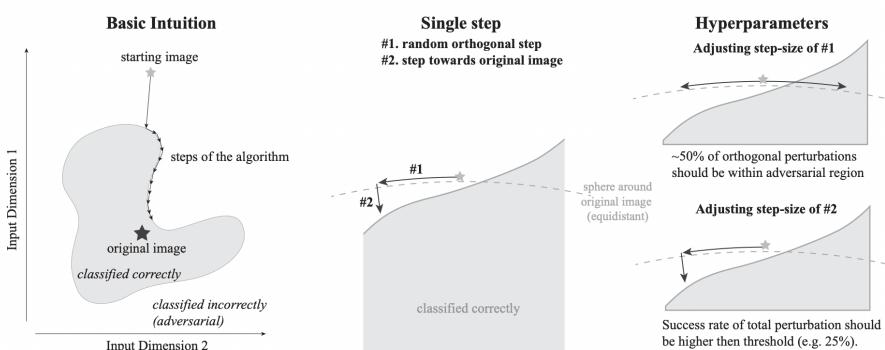
این حمله یک حمله جعبه سیاه مبتنی بر مرز است. این نوع حمله در واقع سخت ترین نوع حمله می‌باشد.

این حمله علیرغم اینکه جعبه سیاه است ولی اندازه نویزی که می‌سازد برای فریب مدل‌های یادگیری ماشین می‌تواند با حمله‌های جعبه سفید رقابت کند. تنظیم ابرپارامترها در این نوع حمله به صورت پویا

انجام می‌شود. این حمله هم می‌تواند به صورت هدفمند برنامه ریزی شود و هم به صورت غیرهدفمند. در این نوع حمله دو تا ابرپارامتر δ و ϵ وجود دارد. مقدار δ نویزی است که هر مرحله به خروجی مرحله قبل اضافه می‌شود و مقدار ϵ برابر است با اندازه‌ای که به سمت عکس پس از نگاشت حرکت خواهیم کرد. اگر بخواهیم این حمله را به صورت غیرهدفمند انجام دهیم کافی است از یک نویز با اندازه مشخص شروع کرده و گام‌های زیر را تا همگرایی برداریم [۲۳].

۱. مقدار δ را اضافه می‌کنیم.
۲. حاصل را بر روی کره‌ای به مرکز عکس واقعی و شعاع اندازه مشخص نگاشت می‌کنیم. عملیات نگاشت به این دلیل انجام می‌شود که می‌خواهیم فاصله نمونه‌های تخاصمی تا نویز ثابت بماند.
۳. پس از آن به اندازه ϵ به سمت عکس واقعی حرکت می‌کنیم. اما این مقدار نباید به حدی باشد که نمونه تخاصمی از دست برود.
۴. تا مرحله‌ای مشخص گام‌های ۱-۲-۳ را تکرار می‌کنیم.

در حمله هدفمند به جای نویز از عکسی در کلاس مدنظر شروع خواهیم کرد.



شکل ۲-۸: مراحل الگوریتم حمله مرز تصمیم [۲۳]

۲-۱۰-۲ حملات همگانی

دسته‌ای از حملات هستند که به دنبال پیدا کردن نویزی می‌باشند که با اضافه کردن آن به هر تصویری، مدل احتمال انتساب آن نمونه تخاصمی به کلاس مطلوب کمتر از کلاسی دیگر شود.



شکل ۲-۹: بررسی نتایج بدست آمده در دو حالت هدفمند و غیرهدفمند و تعداد کوئری های مورد نیاز

روابط مربوط به این نوع حملات در زیر آورده شده است:

$$\tilde{x} = x + r \quad (7-2)$$

$$\|r\|_p \leq \zeta \quad (8-2)$$

$$pr(\hat{c}(\tilde{x})) \neq \hat{c}(x)) \geq 1 - \delta \quad (9-2)$$

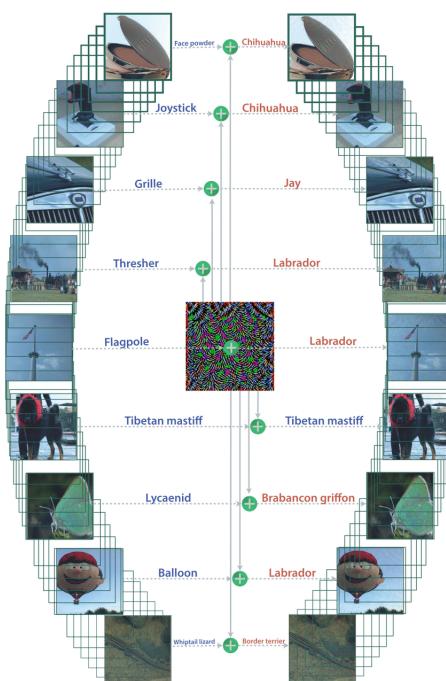
$$\hat{c}(x) = arg \max_k S_k(x) \quad (10-2)$$

همان طور که در شکل زیر مشاهده می کنید تنها یک نویز هدفمند باعث می شود مدل در بسیاری از تصاویر دچار اشتباه شود [۲۷].

۳-۲ الگوریتم های یادگیری تخاصمی

۱-۳-۲ ادبیات موضوع

برای بررسی این الگوریتم ها یک وظیفه دسته بندی را در نظر می گیریم. ورودی مسئله شامل زوج های (x, y) از توزیع D است که x ورودی و y برچسب متناظر با آنرا مشخص می کند. هدف الگوریتم یادگیری در حالت عادی، پیدا کردن پارامتر های θ برای مدل F است به گونه ای که امید ریاضی تابع



شکل ۲-۱۰: نمونه‌ای از حمله همگانی انجام شده [۲۶]

هزینه $E_{(x,y) \sim D}[\text{Loss}(F(x), y|\theta)]$ کمینه شود. این رویکرد، همان الگوریتم کمینه سازی خطر تجربی است که مقاومت تخاصمی بالایی ندارد، زیرا الگوریتم آموزش هیچ توجهی به نمونه‌های خارج از توزیع اولیه داده‌ها و به خصوص نمونه‌های تخاصمی ندارد و هیچ تضمینی برای کم بودن خطا بر روی آن‌ها ندارد.

۲-۳-۲ یادگیری تخاصمی

این روش که به عنوان یکی از اولین روش‌های یادگیری تخاصمی معرفی شده است، مسئله نمونه‌های تخاصمی را از دیدگاه بهینه‌سازی بررسی می‌کند و سعی می‌کند با همین چهارچوب راه حلی برای آن ارائه کند. در این روش، مجموعه Δ به عنوان مجموعه تمام انحراف‌های مجاز (با اندازه محدود) در نظر گرفته می‌شود. با این تعریف، عبارتی که باید توسط الگوریتم آموزش بهینه شود، به رابطه زیر تغییر پیدا می‌کند.

$$\min_{\theta} \{E_{(x,y) \sim D} [\max_{\delta \in \Delta} Loss(x + \delta, y, \theta)]\} \quad (11-2)$$

عبارت فوق یک مسئله بهینه‌سازی دو مرحله ای است که در مرحله اول سعی می‌شود نمونه‌های تخاصمی ای ایجاد کرد که بیشینه هزینه را ایجاد کنند و در مرحله دوم سعی می‌شود مدل نسبت به این نمونه‌ها مقاوم شود.

در این الگوریتم در هر مرحله از بهینه‌سازی به جای استفاده از ورودی تمیز x ، با استفاده از حمله نزول گرادیان افکنده شده نمونه‌های تخاصمی x_{adv} تولید می‌شوند و مدل با مجموعه داده (x_{adv}, y) آموخته می‌بیند.

در این الگوریتم از روش نزول گرادیان دو بار استفاده می‌شود. یک بار برای بیشینه سازی داخلی و یک بار برای کمینه سازی خارجی. به همین دلیل زمان مورد نیاز برای آموخته مدل متناسب با تعداد تکرارهای بهینه‌سازی داخلی افزایش می‌یابد. برای مثال در بهینه‌سازی داخلی، الگوریتم نزول گرادیان افکنده معمولاً حدود ۱۰ مرتبه تکرار می‌شود که باعث می‌شود زمان اجرای الگوریتم ۱۰ برابر شود.

این روش به عنوان پایه ای برای اکثر الگوریتم‌های یادگیری مقاوم استفاده می‌شود. به طوری که تقریباً تمام روش‌هایی که درستی مقاوم بودن آن‌ها ثابت شده‌است، گسترشی از این روش می‌باشد [۱۶].

۳-۳-۲ نقطیر دفاعی

یکی از دفاع‌های شناخته شده در بحث یادگیری تخاصمی نقطیر دفاعی می‌باشد، این دفاع نمونه‌های تخاصمی ایجاد شده بر پایه گرادیان را از کار می‌اندازد و مدل را نسبت به این نوع حملات مقاوم می‌کند. در این نوع دفاع ما دو از دو تا شبکه عمیق استفاده می‌کنیم. یکی شبکه معلم^{۱۲} و یکی شبکه دانش آموز^{۱۳} است. شبکه معلم را با استفاده از Softmax همراه با درجه حرارت T اجرا می‌کنیم، پس از اینکه این شبکه به خوبی آموخته شده باشد می‌بایست خروجی این شبکه را به عنوان خروجی مطلوب شبکه دانش آموز در نظر گرفت و سپس شبکه دانش آموز را با Softmax معمولی اجرا کرد.

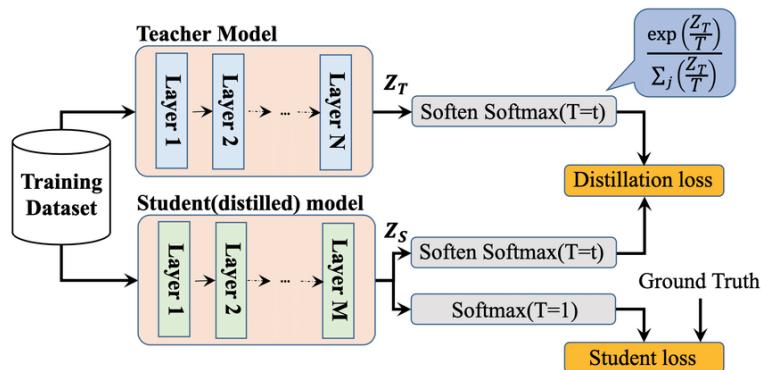
در این نوع از دفاع گرادیان انتقالی ناپدید می‌شود و اندازه نویزی که به تصویر اضافه می‌شود معمولاً بسیار کم است به نحوی که قادر به تولید نمونه تخاصمی نمی‌باشد [۲۸].

¹²Teacher

¹³Student

$$F(X) = \left[\frac{e^{z_i(X)/T}}{\sum_{l=0}^{N-1} e^{z_l(X)/T}} \right]_{i \in 0..N-1}$$

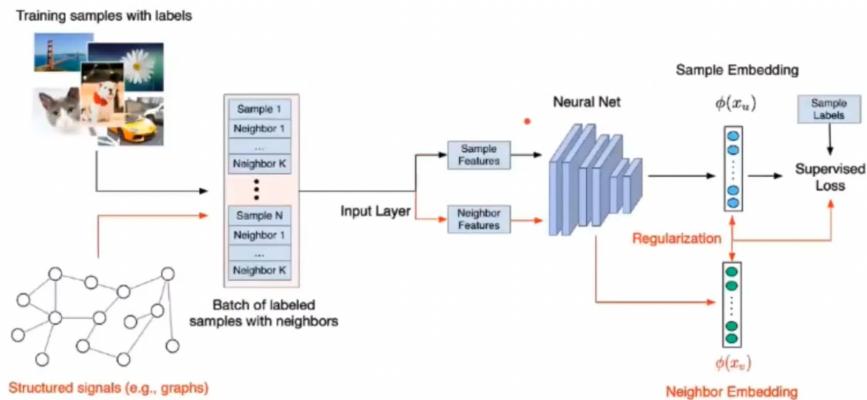
شکل ۲-۱۱: رابطه استفاده شده برای آموزش شبکه معلم [۲۸]



شکل ۲-۱۲: نحوه اجرای الگوریتم تقطیر دفاعی [۲۹]

۲-۳-۴ یادگیری عصبی ساختارمند

در این روش از دسته روش‌های یادگیری مقاوم، علاوه بر مجموعه داده‌هایی که به شبکه تزریق می‌کنیم، یک سری داده‌های تخاصمی ساختگی تحت عنوان همسایه برای بقیه داده‌ها در نظر می‌گیریم. در این شبکه علاوه بر تابع هزینه معمولی که برای شبکه در نظر می‌گیریم یک تابع هزینه دیگر به آن اضافه می‌کنیم که نقش آن نزدیک کردن بردارهای قبل از Softmax برای داده‌های در همسایگی مجاور است. می‌توان اینطور تصور کرد که در این روش یادگیری علاوه بر داده‌های اصلی داده‌هایی با نویز کم در اطراف داده‌های اصلی را نیز یاد می‌گیریم [۳۰].



شکل ۲-۱۳: نحوه کار الگوریتم یادگیری عصبی ساختارمند [۳۰]

TRADES ۵-۳-۲

در این روش نیز از نزول گرادیان استفاده می‌شود، تفاوت این روش، نحوه متفاوت بدست آوردن نمونه‌های تخاصمی و استفاده از تابع هزینه‌ای متفاوت است که باعث بهبود روند یادگیری تخاصمی و دقیق آن می‌شود. مدل بدست آمده در این روش از حل بهینه‌سازی زیر بدست می‌آید.

$$\min_F \{Loss(F(x), y) + \max_{x' \in \mathcal{B}(x, \epsilon)} Loss(F(x), F(x'))/\lambda\} \quad (12-2)$$

در رابطه بالا، $\mathcal{B}(x, \epsilon)$ مجموعه تمام نقاطی است که فاصله آن‌ها از x کمتر مساوی ϵ است. همچنین احتمال نسبت داده شده به هر کلاس توسط مدل را نشان می‌دهد و $Loss$ تابع هزینه مورد استفاده $F(x)$ را نشان می‌دهد (معمولاً بی نظمی تقاطعی^{۱۴} در نظر گرفته می‌شود). در واقع در این روش، نمونه‌های تخاصمی نقاطی از همسایگی ورودی‌اند که پیش‌بینی مدل برای آن‌ها بیشترین تفاوت را با ورودی دارد. نشان داده شده است که این روش باعث می‌شود مرز تصمیم‌گیری از داده‌های ورودی دور شود. این کار باعث می‌شود که اندازه نویز مورد نیاز برای تغییر تصمیم مدل افزایش یابد که در نهایت منجر به مقاوم شدن مدل می‌شود [۳۱].

¹⁴Entropy Cross

۶-۳-۲ روش‌های اثبات پذیر مقاوم

دسته‌ای از روش‌ها^{۳۴، ۳۳، ۳۲} [۳۲] ارائه شده اند که با استفاده از کران گذاری بر روی میزان تغییر هرنورون نسبت به تغییرات ورودی و کمینه کردن این تغییرات، سعی در مقاوم سازی مدل دارند. از نقاط مثبت این روش‌ها، قابل اثبات بودن مقاومت آن‌ها نسبت به حمله‌هایی است که در شرایط کران صدق می‌کنند. اما از آنجایی که این روش‌ها نیاز به محاسبه‌ی کران برای تمام نورون‌های شبکه دارند، استفاده از این روش‌ها تاکنون محدود به مجموعه داده‌ها و شبکه‌های بسیار کوچک بوده است.

تنها استثنای تاکنون، روش‌های هموارسازی تصادفی^{۱۵} بوده است که نتوانسته است بر روی مجموعه داده‌های بسیار بزرگ مورد استفاده قرار بگیرد. این روش [۳۵] می‌تواند از روی هردسته‌بند دلخواهی یک نسخه هموار بسازد و نشان می‌دهد که این نسخه هموار، نسبت به حملات تخاصمی مقاوم‌اند.

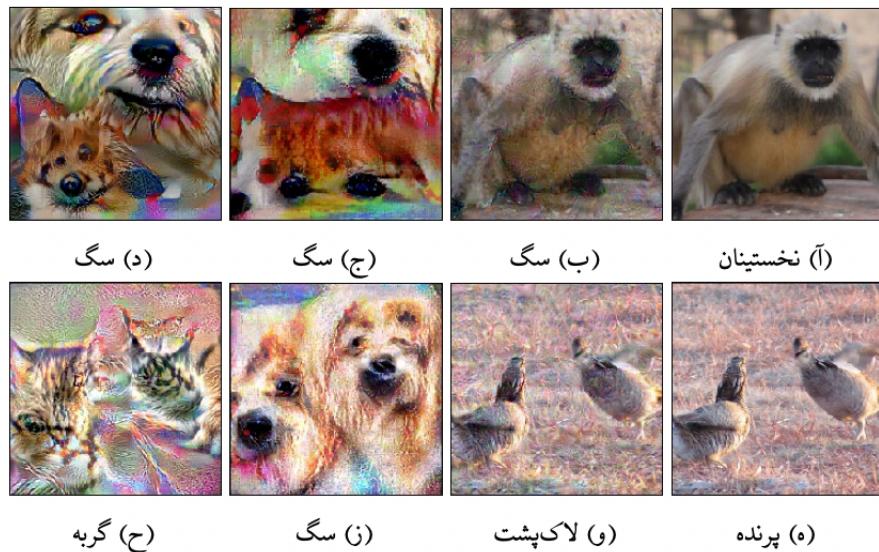
۴-۲ تاثیر یادگیری مقاوم

شاید این سوال مطرح شود که آیا مقاوم سازی مدل‌های یادگیری ماشین می‌تواند عملکرد آن‌ها را بر روی داده‌ها تضعیف کند؟

استفاده از یادگیری مقاوم، علاوه بر مقاوم سازی مدل بر روی حمله تخاصمی مورد نظر، نتایج دیگری را نیز به همراه دارد. عموماً بررسی‌ها نشان داده‌اند که مقاومت بیشتر تخاصمی، کاهش دقت در برابر نمونه‌های تمیز را به همراه دارد [۳۶]. اما برخی پژوهش‌های دیگر نیز ثابت کرده‌اند که با فرض جدایی پذیر بودن داده‌ها (که این فرض در مجموعه داده‌های طبیعی برقرار است) می‌توان مدلی ساخت که هم مقاوم و هم دقیق باشد [۳۷]. همچنین نشان داده شده مقاومت در برابر حمله‌های تخاصمی ارتباط مستقیم با مقاومت در برابر نویز طبیعی و گاووسی دارد. افزایش هرکدام افزایش دیگری را به همراه دارد [۳۸].

یکی از مهم ترین فواید استفاده از یادگیری مقاوم تفسیرپذیر شدن مدل است. گرادیان ورودی نسبت به خروجی مدل، با درک انسان منطبق‌تر می‌شود که باعث می‌شود تفسیرهای بدست آمده، که عموماً وابسته به گرادیان هستند، منطقی‌تر به نظر برسند. همچنین نشان داده شده است که ویژگی‌هایی که مدل‌های مقاوم یاد می‌گیرد، تفاوت بسیاری با ویژگی‌های مدل عادی دارد [۴۰].

^{۱۵}Randomized Smoothing



شکل ۲-۱۴: ویژگی‌های مورد توجه مدل‌های عادی و مقاوم. ستون اول از راست تصویر اصلی را نشان می‌دهد. ستون بعدی به ترتیب کمترین میزان انحراف مورد نیاز برای فریب دادن مدل‌های استاندارد، مقاوم به نویز ℓ_2 و مقاوم به نویز ℓ_∞ را نشان می‌دهند. برچسب نسبت داده شده توسط هر مدل، زیر هر عکس قرار دارد [۳۹].

همان‌طور که در شکل ۲-۱۴ دیده می‌شود، برای فریب دادن مدل استاندارد به نویز بسیار کوچکی نیاز است که برای چشم غیرمسلح به سختی قابل دیدن است. از طرف دیگر، برای عوض کردن نظر مدل‌های مقاوم، حمله نیاز دارد تا ویژگی‌های تصویر را به کلی تغییر دهد. با وجود تغییری با این میزان در ورودی، تغییر نظر مدل منطقی به نظر می‌رسد. این گواهی بر تفسیرپذیر بودن ویژگی‌های یاد گرفته شده توسط مدل‌های مقاوم است.

۲-۵ روش‌های حذف اثر دسته‌ای

در این بخش، تعدادی از روش‌هایی که سعی در حذف اثر دسته‌ای دارند را بررسی می‌کنیم. این روش‌ها عموماً مبتنی بر پیدا کردن یک نمایش از داده‌اند که ویژگی‌های دسته‌ای در آن وجود نداشته باشد یا تا حد امکان کم شده باشد.

۲-۵-۱ یادگیری مقابله‌ای عمیق در معیارهای زیستی

این روش [۴۱] در حالت کلی سعی می‌کند با استفاده از یک شبکه رمزگذار^{۱۶} نمایشی برای داده‌ها پیدا کند، به نحوی که این نمایش برای وظیفه پایین دستی^{۱۷} مفید باشد اما تا حد امکان اطلاعات کمی از عوامل ایجاد اثر دسته‌ای مانند جلسه ضبط داده داشته باشد.

در این مقاله داده‌هایی از افراد مختلف در جلسه‌های متفاوت ضبط شده‌است و هدف پیدا کردن یک اثر انگشت^{۱۸} برای افراد است. هر داده به شکل (X_i, s_i, r_i) است که در آن X_i سیگنال ضبط شده از فرد s_i در جلسه r_i است. حال سه شبکه زیر را در نظر می‌گیریم. شبکه $g(X; \theta)$ از روی ورودی، نمایشی مانند γ می‌سازد. شبکه $p(l, \phi)$ از روی این نمایش سعی در دسته‌بندی فرد یا حدس زدن s دارد. همچنین شبکه $p(r|g(X; \theta); \phi)$ سعی در حدس جلسه r دارد. اگر تابع هزینه را به شکل زیر معرفی کنیم:

$$\min_{\theta, \gamma} \max_{\phi} E[-\log q(s|g(X; \theta); \gamma) + \lambda \log p(r|g(X; \theta); \phi)] \quad (13-2)$$

مدل به هدف مورد نظر می‌رسد.

۲-۵-۲ شبکه‌های باقی‌مانده‌ای یکسان‌کننده توزیع

در این روش [۴۲] فرض شده‌است که داده‌ها از دو توزیع D_1 و D_2 آمده‌اند. همچنین فرض شده تابع ψ وجود دارد که اگر $X \sim D_1 \sim D_2$ آنگاه $\psi(X) \sim \psi$. همچنین فرض شده‌است که تابع ψ نزدیک به تابع همانی باشد (علت این فرض این است که می‌خواهیم این تابع دقت دسته‌بندی اصلی را تغییر ندهد)، برای حذف اثر دسته‌ای، این روش سعی می‌کند تابع ψ را یاد بگیرد. برای این کار از شبکه ResNet استفاده می‌کند که فرمی مشابه فرض مدل دارد [۴۳]. تابع هزینه مدل، بیشینه میانگین اختلاف^{۱۹} است که به نوعی فاصله بین دو توزیع را نشان می‌دهد.

$$MMD(F, p, q) \equiv \sup_{f \in F} (E_{x \sim p} f(x) - E_{x \sim q} f(x)) \quad (14-2)$$

رابطه ۲-۱۴ نحوه محاسبه معیار بیشینه میانگین اختلاف بین دو توزیع p و q بر روی خانواده توابع F را نشان می‌دهد. این معیار، برابر با کران بالای اختلاف امید ریاضی توابع داخل F است. این روش، با

¹⁶Encoder

¹⁷Downstream Task

¹⁸Fingerprint

¹⁹Maximum Mean Discrepancy

استفاده از رابطه ψ زیر نگاشت ψ را پیدا می‌کند که توزیع دسته دوم را به توزیع دسته اول تبدیل می‌کند.

$$\psi = \arg \min_{\psi} MMD(\{\psi(x_1), \dots, \psi(x_n)\}, \{y_1, \dots, y_m\}) \quad (15-2)$$

۳-۵-۲ حذف اثر دسته‌ای به کمک رمزگذاری مستقل از دسته

در این روش به ازای دو دسته داده X_A و X_B سعی می‌شود نمایشی تولید شود که اطلاعاتی از دسته نداشته باشد و فقط شامل اطلاعات زیستی باشد. برای این کار از یک کدگذار E استفاده می‌شود. برای هر دسته هم از یک رمزگشا^{۲۰} استفاده می‌شود تا بتواند نمایش آن دسته را به حالت اولیه بازگرداند. بنابراین، اگر نمایش داده‌ها مستقل از دسته آن‌ها باشد، نیاز است که اطلاعات مربوط به دسته، در رمزگشای آن دسته ثبت شود. برای این که نمایش مستقل از دسته باشد نیز از یک شبکه متخصص دسته‌بند بر روی نمایش استفاده می‌شود که کدگذار سعی در فریب دادن آنرا دارد [۴۴].

²⁰Decoder

فصل ۳

روش پیشنهادی

مجموعه داده‌هایی که اثر دسته‌ای در آن‌ها وجود دارد، از طریق نمونه‌برداری‌های متعدد از دسته‌های متفاوت بدست آمده اند که شرایط محیطی و اندازه گیری لزوماً در همه آن‌ها یکسان نبوده است، به همین دلیل علاوه بر ویژگی‌های مورد نظر، ویژگی‌های مختص هر دسته به نمونه‌ها اضافه شده است. وجود چنین ویژگی‌هایی در مجموعه داده، باعث می‌شود که مدل‌های یادگیری، از هویت دسته برای تشخیص برچسب مورد نظر استفاده کنند. در حالیکه چنین ویژگی‌هایی هیچ ارتباط معناداری با آن برچسب ندارند. ما با استفاده از یادگیری تخصصی، روشی برای کم کردن تاثیر ویژگی‌های بوجود آمده توسط اثر دسته‌ای ارائه خواهیم کرد. در این روش سعی می‌شود نمایشی از داده‌های اولیه به دست بیاید که همبستگی ویژگی‌های حاصل از اثر دسته‌ای و برچسب مورد نظر در داده‌های اصلی تا حد امکان کاهش یابد. بنابراین، مدل یادگیری نمی‌تواند از این ویژگی‌ها برای آموزش و تمایز استفاده نماید. علاوه بر این، نمایش به دست آمده در همان فضای اولیه داده‌ها قرار دارد و ابعاد یکسانی با داده‌های اولیه دارد. به همین علت می‌توان روش ارائه شده را یک روش داده افزایی^۱ تلقی کرد که به نوعی با معرفی داده‌هایی که ویژگی‌های پایه‌ای یکسان و ویژگی‌های دسته‌ای متفاوت دارند، سعی دارد مدل را از یاد گرفتن ویژگی‌های دسته‌ای بازدارد [۳۹].

در این روش فرض می‌شود هر نمونه ورودی هم شامل ویژگی‌های پایه‌ای و هم ویژگی‌های دسته‌ای است. به صورت واضح‌تر فرض می‌کنیم هر نمونه x طبق رابطه

$$x = u + v \quad (1-3)$$

¹Data Augmentation

از جمع ویژگی‌های پایه‌ای u و دسته‌ای v ساخته شده است. همچنین برچسب‌های y و z که به ترتیب مربوط به وظیفه دسته‌بندی خواسته مسئله و دسته‌ای که نمونه x از آن آمده است، موجود هستند. موجود بودن برچسب z فرض بسیار ساده و معقولی می‌باشد، زیرا این برچسب می‌تواند شماره جلسه^۲ جمع آوری داده، هویت فرد^۳ مورد آزمایش، شماره ظرف^۴ و غیره باشد که در عموم مجموعه داده‌های مربوط به مسائل پژوهشی موجود می‌باشند. همچنین فرض می‌شود توابع f و g وجود دارند به‌طوری‌که $y = f(x)$ و $z = g(x)$. تابع f نشان دهنده برچسب وظیفه دسته‌بندی مورد نظر است که طبق تعریف، وابسته به ویژگی‌های پایه‌ای موجود در نمونه u است. به صورت مشابه تابع g تنها به v وابسته است [۳۹].

الگوریتم ۱ الگوریتم آموزش مدل مقاوم

ورودی: (x, y, z)

ورودی: F, θ_F, G, θ_G

ورودی: \mathcal{L}

ورودی: $\epsilon, \alpha_1, \alpha_2$

$$\theta_G \leftarrow \theta_G - \alpha_1 \nabla_{\theta_G} \mathcal{L}(G(x, z | \theta_G)) : ۱$$

$$x' \leftarrow \arg \max_{w \in \mathcal{B}(x, \epsilon)} \nabla_{\theta_G} \mathcal{L}(G(w), z | \theta_G) : ۲$$

$$\theta_F \leftarrow \theta_F - \alpha_2 \nabla_{\theta_F} \mathcal{L}(F(x', y | \theta_F)) : ۳$$

سپس از دو مدل F و G برای تخمین توابع f و g استفاده می‌کنیم. این دو مدل می‌توانند متفاوت از یکدیگر باشند اما از آنجایی که ابعاد ورودی‌شان یکسان است، ما آن‌ها را یکسان در نظر می‌گیریم. فرض می‌کنیم نمونه‌های ورودی به صورت سه تایی (x, y, z) هستند به‌طوری‌که x سیگنال ورودی، y برچسب وظیفه مورد نظر و z برچسب دسته‌ای است که x به آن تعلق دارد را نشان می‌دهد. به ازای هر سه تایی ورودی، ابتدا با استفاده از x و برچسب z مدل G را آموزش می‌دهیم. سپس یک نمونه تخاصمی^۵ از روی x با استفاده از رابطه‌ی

$$x' = \arg \max_{w \in \mathcal{B}(x, \epsilon)} \mathcal{L}(G(w), z | \theta_G) \quad (۲-۳)$$

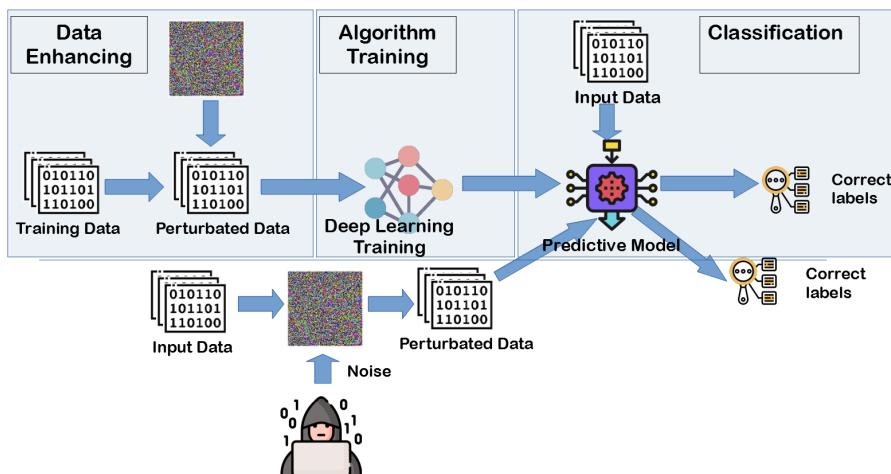
می‌سازیم. در حالتی که مدل G بتواند دسته‌ای که داده از آن آمده است را تشخیص دهد، نمونه x' که از رابطه‌ی بالا بدست می‌آید، با حذف ویژگی‌های موثر در دسته‌بندی، سعی در به خط انداختن مدل خواهد

²Session

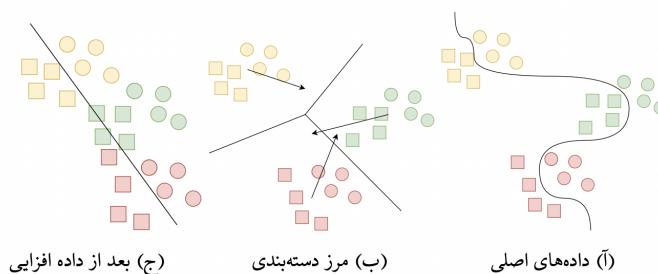
³Identity

⁴Plate

داشت. بنابراین، می‌توان فرض کرد که تفاوت x' با x در ویژگی‌های مربوط به دسته است. زیرا تنها این ویژگی‌ها هستند که می‌توانند در تصمیم‌گیری مدل G موثر باشند. سپس با استفاده از نمونه تخصصی بدست آمده مدل F را با استفاده از برچسب y آموزش می‌دهیم. خلاصه روش آموزش مدل‌هارا می‌توان در الگوریتم ۱ مشاهده کرد. ورودی‌های این الگوریتم شامل سه تایی ورودی (x, y, z) ، مدل‌های F و G که به ترتیب توسط θ_F و θ_G پارامتریزه شده‌اند،تابع هزینه \mathcal{L} ، نرخ‌های یادگیری^۵ α_1 و α_2 و بودجه انحراف ϵ می‌باشد [۳۹].



شکل ۳-۱: شکل مربوط به الگوریتم حذف اثر دسته‌ای. ورودی هر دو شبکه یکسان می‌باشد و نمونه تخصصی از شبکه مربوط به دسته‌بندی کلاس‌ها ایجاد می‌شود [۴۵].



شکل ۳-۲: تاثیر روش پیشنهادی. دایره و مربع، دسته‌های وظیفه دسته‌بندی اصلی و رنگ‌ها، اثرات دسته‌ای را نشان می‌دهد [۳۹].

اثر روش پیشنهادی در شکل ۳-۲ قابل مشاهده است. داده‌های اصلی تحت تاثیر اثر دسته‌ای، توزیع پیچیده‌ای دارند و مدل ممکن است دچار بیش برازش شود و مرز دسته‌بندی پیچیده‌ای را یاد بگیرد که

⁵Learning Rate

داده‌های آموزش را به خوبی جدا می‌کند. اما بر روی داده‌های آزمون که ممکن است از دسته‌ای دیگر باشند، بد عمل کند. روش ارائه شده، سعی می‌کند، دسته‌ها را به هم نزدیک تر کند. در واقع هر حمله تخاصمی بر روی برچسب دسته، سعی دارد نمونه مورد نظر را از نزدیکترین مرز دسته‌بندی رد کند تا مدل را فریب دهد. بعد از حمله، داده‌های دسته‌های مختلف به همدیگر نزدیکتر می‌شوند و ویژگی‌های دسته برای یادگیری به منظور دسته‌بندی بر اساس برچسب اصلی، کم ارزش تر می‌شوند. به همین دلیل انتظار داریم که مدل از انها صرف نظر کند. ممکن است بعد از حمله، داده‌ها به گونه‌ای قرار گیرند که کاملاً جدا شدنی نباشند و دقت دسته‌بندی بر روی داده‌های آموزش قدری کاهش یابد [۳۹].

فصل ۴

ارزیابی و نتایج

۱ - ۴ مجموعه داده‌ها

برای ارزیابی روش ارائه شده، از دو مجموعه داده SA و DEAP استفاده می‌کنیم. دلیل انتخاب این دو مجموعه داده، وجود تفاوت‌های حاصل از اثر دسته‌ای است، وجود آزمایش‌های متعدد یکسان و وجود افراد آزمایش شونده مختلف باعث تغییر تابع توزیع داده‌ها می‌شود. جدول ۴ - ۱ خلاصه‌ای از ویژگی‌های هر مجموعه داده را نشان می‌دهد.

جدول ۴-۱: ویژگی‌های مجموعه داده

DEAP	SA	ویژگی
۳۲	۱۹	تعداد افراد مورد آزمایش
۱	۲	تعداد حالت‌های آزمایش
۴۰	۱۲۵	تعداد آزمایش‌های هر حالت
۳۲	۲۷۰	تعداد حسگرها
۳۸۴ * ۳۲	۱۲۰۰ * ۱۱۶	ابعاد ورودی
۲	۲	تعداد دسته‌های پیش‌بینی

مجموعه داده "میرایی حسی"^۱ که به صورت سیگنال‌های MEG^۲ ضبط شده است، حاوی سیگنال‌های مغزی ۱۹ فرد مورد آزمایش است. هر آزمایش، در یکی از دو حالت فعال^۳ و غیرفعال^۴ انجام می‌شود. در حالت غیرفعال یک صدا در فواصل منظم برای فرد پخش می‌شود. در حالت فعال، از فرد خواسته می‌شود که دکمه‌ای را هر ۳ ثانیه فشار دهد و صدا در همان لحظه برای وی پخش می‌شود. مدت زمان ضبط سیگنال در این مجموعه داده ۴ ثانیه است که به صورت متقاضن حول زمان پخش صدا قرار دارد. جزئیات مجموعه داده را می‌توان در منبع ۳۲ که مجموعه داده را معرفی کرده است، یافت. در این مجموعه داده، هدف تشخیص نوع حالت آزمایش (فعال - غیرفعال) از روی سیگنال ورودی خواهد بود [۳۹، ۴۶].

مجموعه داده "بررسی احساسات توسط سیگنال‌های تنکردنی"^۵ که به صورت سیگنال‌های EEG^۶ و تعدادی سیگنال تنکردنی دیگر ضبط شده است، از ۳۲ فرد استخراج شده است. به هر فرد، تعدادی نماهنگ^۷ نشان داده می‌شود و از آن‌ها خواسته می‌شود پس از مشاهده‌ی هر نماهنگ، به هر یک از احساسات برانگیختگی^۸، ظرفیت^۹ و غلبه^{۱۰} امتیازی از ۱ (کم) تا ۹ (زیاد) نسبت دهند. دو ویژگی برانگیختگی و ظرفیت از دیدگاه ۳۳ می‌تواند روش خوبی برای عددی ساختن و گسترش سازی احساسات

¹Sensory Attenuation²Magnetoencephalogram³Active⁴Passive⁵Database For Emotion Analysis Using Physiological Signals⁶Electroencephalogram⁷Music Video⁸Arousal⁹Valence¹⁰Dominance

باشد. برای همین در این کار، ما از این دو ویژگی استفاده می‌کنیم. علاوه بر این، مطابق کارهای مشابه انجام شده بر روی این مجموعه داده، با استفاده از آستانه^{۱۱} ۵، این ویژگی‌ها را به ویژگی‌هایی دودویی^{۱۲} بالا^{۱۳} و پایین^{۱۴} تبدیل می‌کند. مدت زمان ضبط سیگنال در این مجموعه داده ۱ دقیقه و ۳ ثانیه است که ۳ ثانیه اول آن قبل از پخش نماهنگ بوده و ما از آن صرف نظر می‌کنیم. همچنین، مطابق کارهای مشابه انجام شده، بازه‌ی ۱ دقیقه ای ضبط شده به بازه‌های ۳ ثانیه‌ای به همراه اشتراک شکسته می‌شود. بنابراین، از هر نمونه ۱ دقیقه ای، ۵۸ نمونه ۳ ثانیه‌ای بدست خواهیم آورد. جزئیات مربوط به این مجموعه داده را می‌توان در منبع^{۱۵} ۳۴ مشاهده کرد. همچنین از میان داده‌ها، تنها داده‌هایی را انتخاب می‌کنیم که ویژگی ظرفیت آن‌ها پایین بوده است و برچسب مورد نظر دسته‌بندی را ویژگی برانگیختگی در نظر می‌گیریم [۴۷، ۳۹].

۱-۱-۱ داده‌های مسئله

مگنتوانسفالوگرافی

مغناطیس نگاری مغزی یا مگنتوانسفالوگرافی^{۱۶} یک فن تصویربرداری عصبی کاربردی^{۱۷} برای نقشه برداری فعالیت مغز با ثبت میدان‌های مغناطیسی تولید شده توسط جریان‌های الکتریکی که به طور طبیعی در مغز رخ می‌دهد، با استفاده از مغناطیس سنج‌های بسیار حساس است. آرایه‌های SQUID در حال حاضر رایج ترین مغناطیس سنج‌ها هستند. کاربردهای MEG شامل تحقیقات پایه‌ای در فرایندهای ادراکی و شناختی مغز، محلی سازی مناطق آسیب شناسی قبل از جراحی، تعیین عملکرد بخش‌های مختلف مغز و نوروفیدبک^{۱۸} است. این را می‌توان در یک محیط بالینی برای یافتن مکان‌های ناهنجاری و همچنین در محیط‌های آزمایشی برای اندازه‌گیری ساده فعالیت مغز اعمال و استفاده کرد [۴۸].

الکتروانسفالوگرافی

مغز انسان متشكل از میلیون‌ها نورون^{۱۹} است که نقش مهمی در کنترل رفتار بدن انسان با توجه به محرک‌های حرکتی/حسی داخلی/خارجی دارند. این نورون‌ها به عنوان حامل اطلاعات بین بدن و مغز

¹¹Threshold

¹²Binary

¹³High

¹⁴Low

¹⁵MEG: Magnetoencephalography

¹⁶functional

¹⁷Neurifeedback

¹⁸Neuron

انسان عمل خواهند کرد. درک رفتار شناختی مغز را می‌توان با تجزیه و تحلیل سیگنال‌ها یا تصاویر از مغز انجام داد. رفتار انسان را می‌توان بر حسب حالات حرکتی و حسی مانند حرکت چشم، حرکت لب، به یاد آوردن، توجه کردن، فشردن دست و... مدل کرد. این حالات با فرکانس سیگنال خاصی مرتبط هستند که به درک رفتار عملکردی ساختار پیچیده مغز کمک می‌کند.

الکتروانسفالوگرافی^{۱۹} روشی کارآمد است که به دریافت سیگنال‌های مغزی مربوط به حالات مختلف از سطح پوست سر کمک می‌کند. این سیگنال‌ها به طور کلی بر اساس فرکانس سیگنال از ۰۰ هرتز تا بیش از ۱۰۰ هرتز به عنوان دلتا، تتا، آلفا، بتا و گاما دسته‌بندی می‌شوند [۴۹].

۲-۴ معیارهای ارزیابی

برای ارزیابی^{۲۰} روش پیشنهادی، معیار دقت دسته‌بندی^{۲۱} را در نظر می‌گیریم. همچنین دو راهکار^{۲۲} مختلف برای ارزیابی در نظر می‌گیریم.

۲-۴-۱ ارزیابی روش درهم

در این حالت، ۸۰ درصد نمونه‌ها، به صورت تصادفی و مستقل از برچسب دسته‌شان برای آموزش مدل انتخاب می‌شوند و باقی مانده داده‌ها برای آزمودن استفاده می‌شوند. اثرات دسته‌ای موجود در این حالت تنها به تفاوت‌های جلسه‌های داده‌گیری هر فرد محدود می‌شود.

۲-۴-۲ ارزیابی روش کنارگذاشتن تکی

در این روش، یک فرد به عنوان فرد آزمون انتخاب می‌شود و تمام نمونه‌های آن برای آزمودن مدل و باقی داده‌ها برای آموزش مدل استفاده می‌شود. در این حالت علاوه بر تفاوت جلسه‌های آزمایش فرد، تفاوت‌های ساختار نمونه‌های این فرد با سایر افراد نیز در اثر دسته‌ایتاشیر می‌گذارد. به همین دلیل این روش ارزیابی، روشی سختگیرانه‌تر از روش درهم خواهد بود.

¹⁹EEG

²⁰Validation

²¹Classification Accuracy

²²Strategy

۳-۴ جزئیات پیاده سازی

در انجام آزمایش‌ها، مقدار میانگین و واریانس هر دو مجموعه داده به ترتیب به صفر و یک اصلاح شده است^{۲۳}. در مجموعه داده DEAP تعداد اندکی از داده‌ها قدر مطلق بسیار زیادی بعد از یکه‌سازی واریانس دارند. به همین علت ورودی‌هایی که مقدار قدر مطلق آن‌ها بیشتر از ۶۰۰ بوده است بریده^{۲۴} شده‌اند. تنها ۰٪/۴^{۲۵} کل داده‌ها نیاز به بریده شدن دارند.

برای پیاده سازی مدل‌های F و G در مجموعه داده‌ی DEAP از دو شبکه $ResNet - 18$ استفاده شده است [۴۳]. نمونه‌های مجموعه داده SA دارای ابعاد بسیار بزرگ‌تری نسبت به نمونه‌های DEAP هستند. به همین دلیل امکان استفاده از مدل ResNet وجود ندارد. به همین علت، از یک مدل پیچشی^{۲۵} با عمق ۸ استفاده شده است. جزئیات مربوط به این مدل در جدول زیر آمده است.

²³Data Normalization

²⁴Clip

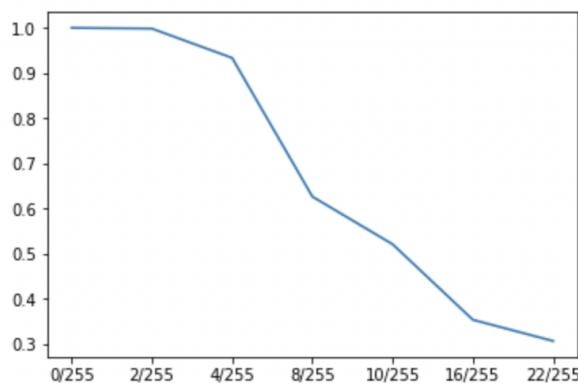
²⁵Convolutional

جدول ۴-۲: معماری شبکه پیچشی استفاده شده

شماره	نوع لایه	ابعاد هسته	طول گام	تابع فعالسازی	توضیحات
۱	پیچشی	۵*۱۱	۱*۲	-	کانال ۱۰
-	هنچارسازی دسته‌ای	-	-	RELU	-
۲	پیچشی	۵*۱۰	۱*۲	-	کانال ۱۰
-	هنچارسازی دسته‌ای	-	-	RELU	-
۳	پیچشی	۵*۱۰	۱*۲	-	کانال ۱۰
-	هنچارسازی دسته‌ای	-	-	RELU	-
۴	پیچشی	۵*۵	۲*۲	-	کانال ۲۰
-	هنچارسازی دسته‌ای	-	-	RELU	-
۵	پیچشی	۵*۵	۲*۲	-	کانال ۲۰
-	هنچارسازی دسته‌ای	-	-	RELU	-
۶	پیچشی	۵*۵	۲*۲	-	کانال ۲۰
-	هنچارسازی دسته‌ای	-	-	RELU	-
-	ادغام پیشینه	۲	-	-	-
-	تماما متصل	۱۴۰۰*۵۱۲	-	Tanh	-
-	Dropout	-	-	-	احتمال ۰.۵
۸	تماما متصل	۵۱۲*۲	-	Softmax	-

برای بهینه سازی تمام مدل‌ها از بهینه ساز^{۲۶} Adam با پارامتر نرخ یادگیری ۰/۰۱ استفاده شده است و یک برنامه ریز^{۲۷} برای کاهش نرخ یادگیری^{۲۸} هر ۲۰ دوره در نظر گرفته شده است. مدل مجموعه داده SA برای ۵۰ دوره^{۲۹} با اندازه دسته^{۳۰} ۶۴ و مدل DEAP برای ۶۰ دوره با اندازه دسته ۱۲۸ آموزش داده شده‌اند.

²⁶Optimizer²⁷Scheduler²⁸Learning Rate²⁹Epoch³⁰Batch Size



شکل ۴-۱: دقت مدل‌های استاندارد نسبت به حمله‌های تخاصمی

برای تولید نمونه‌های تخاصمی، از الگوریتم نزول گرادیان افکنده شده با معیار فاصله ℓ_∞ استفاده می‌کنیم. پارامتر بودجه انحراف (ϵ) را با استفاده از حمله تخاصمی بر روی مدل استاندارد به دست می‌آوریم، برای بودجه‌های انحراف مختلف میزان دقت مدل در تشخیص فرد، در حضور حمله تخاصمی با انحراف مورد نظر را بدست می‌آوریم. ما دنبال کوچکترین انحرافی هستیم که بتواند مدل را تا حد خوبی فریب دهد. شکل ۴-۱ رابطه‌ی اندازه انحراف و میزان فریب دادن مدل را نشان می‌دهد.

۴-۴ ارزیابی دقت

نتایج ارزیابی مجموعه داده SA در جدول ۴-۳ آورده شده است.

جدول ۴-۳: دقت مدل‌ها، مجموعه داده SA

بودجه انحراف ($4 * 10^{-3}$)					-
روش ارزیابی	۱۶	۱۰	۸	۰	
درهم	۰/۹۱۴	۰/۹۳	۰/۸۸۲	۰/۹۰۷	
کنارگذاشته شده	۰/۵۶۱	۰/۶۷	۰/۵۹	۰/۵۵۹	

جدول ۴-۴: دقت مدل‌ها، مجموعه داده DEAP

بودجه انحراف ($4 * 10^{-3}$)					-
روش ارزیابی	۱۶	۱۰	۸	۰	
درهم	۰/۹۵۱	۰/۹۵۵	۰/۹۶	۰/۹۵۵	
کنارگذاشته شده	۰/۵۸۹	۰/۶۲۱	۰/۶	۰/۶	

در این مجموعه، بدون استفاده از روش پیشنهادی (معادل حالت انحراف صفر)، مدل پیچشی در حالت درهم می‌تواند تا ۹۰ درصد دقت داشته باشد. زمانی که بودجه انحراف افزایش می‌یابد، نمونه‌های تخاصمی تولید شده از دسته اولیه فاصله بیشتری می‌گیرند و احتمال دارد دسته جدیدی تولید کنند در نتیجه باید ϵ کمترین مقداری باشد که توزیع‌هارا یکسان کند. بهترین نتیجه برای بودجه انحراف $\frac{1}{255}$ بدست آمده است. در حالت کنارگذاشته شده، به دلیل این که نمونه‌های فرد آزمون به صورت کامل از داده‌های آموزش حذف شده است، آموزش مدل به شدت سخت تر می‌شود. همان طور که دیده می‌شود، بدون استفاده از روش پیشنهادی، مدل قادر به یادگیری نیست و دقت آن با مدلی که به صورت تصادفی پیش‌بینی انجام می‌دهد یکسان است. اما با بکارگیری روش پیشنهادی، دقت به طرز قابل توجهی افزایش پیدا می‌کند. بهترین دقت مربوط به مدل با انحراف $\frac{1}{255}$ است.

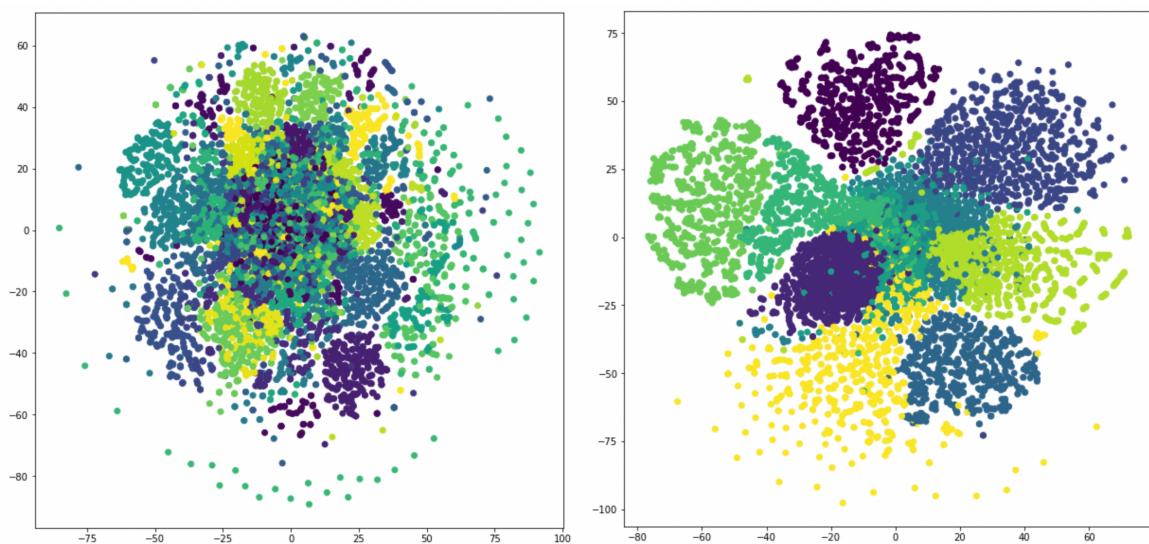
برای مدل با انحراف بیشتر، افت دقت مشاهده می‌شود. به نظر می‌رسد که با افزایش بیش از حد بوجه انحراف، ویژگی‌های مفید دسته‌بندی هم تحت تاثیر قرار می‌گیرند و از داده‌های اصلی حذف می‌شوند. نتایج مجموعه داده DEAP در جدول ۴-۴ آمده است، دقت‌ها برای این مجموعه داده، با توجه به پیچیدگی بیشتر مدل ResNet و همچنین کوچک بودن ابعاد ورودی، بسیار بالاتر از مجموعه داده SA است. همان طور که در جدول دیده می‌شود، روش پیشنهادی توانسته است دقت دسته‌بندی درهم را ۰.۵ درصد افزایش دهد که نشان دهنده اثر بخش بودن حمله تا حدی بوده است. همان طور که مشاهده می‌کنید اگر مقدار بودجه انحراف را از حدی بیشتر کنیم باعث می‌شود ویژگی‌های موثر هم دچار تغییرات

ناخواسته شوند و دقت مدل روی فرایند اصلی دستخوش تغییر شود. در حال کنارگذاشتن یک دسته مدل پیشنهادی توانسته است دقت بر روی داده تست را ۲ درصد افزایش دهد.

۴-۵ ارزیابی کیفی

در این بخش سعی داریم با استفاده از روش، T-SNE تغییرات حاصل از اعمال روش پیشنهادی را بررسی کنیم [۵۰]. این روش که برای کاهش ابعاد داده‌های با بعد بالا معرفی شده است، امکان به تصویر کشیدن نمونه‌های داده‌هارا فراهم می‌کند. این روش سعی می‌کند نمایش با بعد پایین داده‌ها را به گونه‌ای پیدا کند که توزیع فاصله میان جفت نقاط در نمایش اصلی داده‌ها و نمایش با بعد پایین تا حد امکان نزدیک به هم باشند. به این ترتیب نمایش پیدا شده تا حد امکان به داده‌های اصلی سازگار خواهد بود. هر یک از نقاط دارای ۱۳۹۲۰۰ بعد است که در این حالت برای بهینه سازی زمان ابتدا با استفاده از الگوریتم کاهش بعد بوسیله بررسی مولفه‌های اساسی^{۳۱} ابتدا بعد داده‌هارا به ۳۰۰ کاهش می‌دهیم و سپس با استفاده از الگوریتم T-SNE داده‌هارا به فضای دو بعدی نگاشت می‌کنیم تا قابل نمایش باشند. شکل زیر نتیجه این فرایند را نشان می‌دهد.

³¹PCA



شکل ۴-۲: توزیع داده‌ها قبل از اجرای حمله شکل ۴-۳: توزیع داده‌ها پس از اجرای حمله همانطور که در شکل بالا مشاهده می‌شود، در داده‌های اصلی، نمونه‌های هر فرد، ناحیه خاصی از نمودار را اشغال کرده است. در واقع فاصله بین نمونه‌های هر فرد از فاصله بین نمونه‌های دو فرد مختلف عموماً کمتر است. این پدیده به نوعی اثر دسته‌ای‌القا شده توسط هر فرد را نشان می‌دهد که باعث می‌شود نمونه‌های هر دسته در اینجا فرد، توزیعی متفاوت با دسته‌های دیگر داشته باشد. در شکل ب، بعد از اعمال روش پیشنهادی، دیده می‌شود که توزیع داده‌ها در دسته‌های مختلف شباهت بیشتری به هم پیدا کرده اند و تمایز دادن دسته‌های مختلف مانند قبل راحت نیست. این تصویر گواهی به توانایی روش ارائه شده برای حذف اثر دسته‌ای‌بیش‌برازش مدل‌های مختلف می‌باشد.

۴-۶ تاثیر بودجه انحراف بر بیش‌برازش مدل‌ها

اثرات دسته‌ای‌بیا ایجاد تغییرات ناخواسته در توزیع داده‌ها، باعث بوجود آمدن ویژگی‌هایی جدید می‌شوند که ارتباط علی با برچسب دسته‌بندی واقعی ندارند اما همبستگی غیربدیهی با آن دارند. به همین دلیل مدل‌های یادگیری، قسمتی از قدرت پردازشی خود را صرف یادگیری چنین ویژگی‌هایی در داده‌های آموزشی می‌کنند (مسئله اسب باهوش!). اما در داده‌های آزمون چنین همبستگی ممکن است وجود نداشته باشد یا به گونه‌ای دیگر باشد. به همین علت، مدل بر روی داده‌های آموزشی دچار بیش‌برازش

^{۳۲}Overfit

می‌شود و توانایی پیش‌بینی بر روی داده‌های جدید را ندارد. در جدول صفحه بعد دقت آموزش مدل‌ها و میزان تفاوت دقت آموزش و آزمون را برای مدل‌های مختلف آورده‌ایم. همانطور که دیده می‌شود میزان بیش‌برازش و دقت آموزش با دقت آزمون رابطه عکس دارند.

در واقع می‌توان نتیجه گرفت که حمله تخاصمی با حذف ویژگی‌های دسته‌ای یادگیری مدل را سخت‌تر کرده است. اما این سخت‌تر شدن باعث شده است مدل‌ها به ویژگی‌هایی که با برچسب رابطه‌على دارند، توجه بیشتری کنند که حاصل آن در نهایت افزایش دقت بر روی داده‌های دیده نشده آزمون است که همان مطلوب هر مسئله یادگیری در حوزه هوش مصنوعی می‌باشد.

جدول ۴-۵: بیشبرازش مدل‌ها، مجموعه داده SA

روش ارزیابی	انحراف ($10^{-3} * 4$)	دقت آزمون	دقت آموزش	بیشبرازش (آزمون-آموزش)
درهم	۰	۰/۹۰۷	۱	۰/۰۹۳
	۸	۰/۸۸۲	۱	۰/۱۱۸
	۱۰	۰/۹۳	۱	۰/۰۷
	۱۶	۰/۹۱۴	۰/۹۸۵	۰/۰۷۱
کنارگذاشته شده	۰	۰/۰۵۹	۱	۰/۴۴۱
	۸	۰/۰۹	۱	۰/۴۱
	۱۰	۰/۶۷	۱	۰/۳۳
	۱۶	۰/۵۶۱	۰/۹۹۱	۰/۴۳

جدول ۴-۶: بیشبرازش مدل‌ها، مجموعه داده DEAP

روش ارزیابی	انحراف ($10^{-3} * 4$)	دقت آزمون	دقت آموزش	بیشبرازش (آزمون-آموزش)
درهم	۰	۰/۹۵۵	۱	۰/۰۴۵
	۸	۰/۹۶	۱	۰/۰۴
	۱۰	۰/۹۵۵	۱	۰/۰۴۵
	۱۶	۰/۹۱۴	۱	۰/۰۹۶
کنارگذاشته شده	۰	۰/۶	۱	۰/۴
	۸	۰/۶	۱	۰/۴
	۱۰	۰/۶۲۱	۱	۰/۳۷۹
	۱۶	۰/۵۸۹	۱	۰/۴۱۱

فصل ۵

نتیجه‌گیری

به کارگیری مدل‌های یادگیری ماشین در زمینه‌ی داده‌های پزشکی با چالش‌های گوناگونی روبه رو است. مجموعه داده‌های موجود عموماً اندازه کوچکی دارند که آموزش مدل‌های پیچیده‌تر را بر روی آنها دشوارتر می‌سازد. هزینه جمع‌آوری بالای داده یکی از مهم‌ترین عوامل محدود کننده‌ی اندازه این مجموعه داده‌ها می‌باشد. از طرفی دیگر، برچسب‌زنی داده‌ها^۱ نیاز به حضور متخصصان دارد که باعث افزایش هزینه‌های جمع‌آوری داده می‌شود. این چالش‌ها محدود به جمع‌آوری داده نمی‌شود. شرایط آزمایشگاهی متفاوت مانند زمان، دما، افراد مورد آزمایش، خطاهای انسانی و یا حتی خطای دستگاه اندازه‌گیری می‌تواند باعث ایجاد تغییراتی در داده‌های بدست آمده شود. این تغییرات که با نام اثرات دسته‌ای شناخته می‌شوند از چالش‌های عمدۀ یادگیری ماشین در زمینه‌های زیستی هستند.

داده‌ها در دسته‌هایی که از آزمایش‌های متفاوت بدست آمده‌اند توزیع‌های متفاوتی از یکدیگر دارند. این تفاوت توزیع در دسته‌های مختلف، یکی از فرض‌های پایه ای الگوریتم کمینه سازی خطای تجربی را نقض می‌کند. توزیع داده‌ها در زمان آموزش و آزمون یکسان نیستند، به همین علت، آموزش مدل‌های یادگیری عمیق به صورت ساده بر روی این مجموعه داده‌ها با مشکل رو به رو می‌شود.

روش‌های متعددی برای حل این مشکل ارائه شده‌اند. عمدۀ‌ی روش‌ها سعی دارند با یکسان سازی توزیع داده‌ها در دسته‌های مختلف، امکان آموزش مدل به صورت سر راست را فراهم آورند. از ایراداتی که به این روش‌ها وارد است، نیازمندی آنها به دانستن توزیع دسته‌هاست، یکسان سازی توزیع دسته‌های جدید که در زمان آموزش دیده نشده‌اند، همواره سخت‌تر بوده است.

¹Labeling

در این پژوهش، با استفاده از یادگیری تخاصمی، چهارچوبی برای مدل‌سازی اثرات دسته‌ای ارائه کردیم. از دیدگاه یادگیری تخاصمی، می‌توان اثر دسته‌ای را مانند حمله‌ای تخاصمی دید که باعث فریب دادن مدل می‌شود. تغییری در ویژگی‌های نامرتبط با وظیفه اصلی مورد نظر و عموماً با اندازه‌های کوچک، که باعث تغییر پیش‌بینی مدل می‌شوند. سپس با استفاده از این مدل‌سازی، روشی ارائه شد که سعی دارد همبستگی میان ویژگی‌های حاصل از اثرات دسته‌ای و برچسب وظیفه مورد نظر را کاهاش دهد.

این روش سعی در حذف اثر دسته‌ای ندارد، بلکه با تغییر دسته‌ای که داده از آن آمده است در حین آموزش، سعی دارد مدل را از یادگیری ویژگی‌های مختص دسته منصرف کند. به این صورت مدل تمرکز خود را به جای ویژگی‌های مرتبط با دسته، به ویژگی‌های مفید و معنی دار معطوف می‌کند. به همین دلیل در زمان آزمودن مدل، دسته‌های جدید نمی‌توانند مدل را گمراه کنند.

همانطور که در فصل ۴ نشان داده شد، این روش قادر است دقت دسته‌بندی بر روی دو مجموعه داده‌ی سیگنال‌های پزشکی را افزایش دهد. علت اصلی این افزایش دقت، معرفی داده‌های جدید بوده است که مانع بیش‌برازش مدل بر روی دسته‌های دیده شده در زمان آموزش شده است.

۱-۵ کارهای پیش رو

۱-۱-۵ استفاده از حمله‌های متفاوت

در این پژوهش حمله‌های تخاصمی استفاده شده محدود به حملات نزول گرادیان افکنده شده با ∞ ^۱ بوده است. این حملات به دلیل امکان پخش‌کردن انحراف به صورت یکسان در تمام درایه‌های ورودی، عموماً برای تصاویر مورد استفاده قرار می‌گیرد. می‌توان با استفاده از حملاتی که به صورت خاص برای سیگنال‌های مغزی طراحی شده‌اند [۵۱]، اثرات دسته‌ای را در مجموعه داده‌های استفاده شده، بهتر شبیه‌سازی کرد.

روش دیگری معرفی شده است به این صورت که به کمک شبکه‌های مولد^۲ تغییر توزیع ایجاد شده توسط اثرات دسته‌ای را یاد گرفت و نمونه‌های تخاصمی موثرتری تولید کرد [۵۲].

¹Generator

۲-۱-۵ یکسان سازی دسته‌ها با تعداد گام‌های متغیر

در روش ارائه شده، با استفاده از حمله‌های تخاصمی، سعی شده نمونه‌هایی تولید شود که ویژگی‌های مرتبط به دسته در آن عوض شده باشد. در حمله گرادیان در نزول افکنده ممکن است دو مشکل وجود داشته باشد.

۱. ویژگی‌های موثر در عمل دسته‌بندی اصلی با این حمله دچار تغییرات شوند و باعث شود دقت مدل افت کند.

۲. حمله باعث بوجود آمدن دسته‌های جدید شود.

یک راه جلوگیری از بوجود آمدن دسته‌های جدید این است که حمله نزول گرادیان افکنده را هدف دار انجام دهیم. به این صورت که همه داده‌های موجود در دسته‌های مختلف را به یک دسته خاص نگاشت کنیم.

یک راه دیگر برای بهبود نتایج در حالت کنارگذاشتن یک دسته خاص می‌تواند این باشد که داده‌ها را با تعداد گام‌های متغیر منحرف کنیم، این کار باعث می‌شود که داده مدنظر پس از به اشتباه انداختن مدل دسته‌بندی دیگر منحرف نشود و پراکندگی کمتری داشته باشیم.

مراجع

- [1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014.
- [2] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015*, 2015.
- [3] Adversarial fgsm. https://www.tensorflow.org/tutorials/generative/adversarial_fgsm.
- [4] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, and C. Xiao. Robust physical-world attacks on deep learning visual classification. *arXiv*, 2018.
- [5] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter. A general framework for adversarial examples with objectives. *arXiv*, 2019.
- [6] J. Kos and D. Song. Delving into adversarial attacks on deep policies. *arXiv*, 2017.
- [7] N. Carlini and D. Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. *arXiv*, 2018.
- [8] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart. Stealing machine learning models via prediction apis. *arXiv*, pages 601–618, 2016.
- [9] White box and black box attacks. <https://medium.com/onfido-tech/adversarial-attacks-and-defences-for-convolutional-neural-networks-66915ece52e7>, 2018.

- [10] U. Shaham, K. P. Stanton, J. Zhao, H. Li, K. Raddassi, R. Montgomery, and Y. Kluger. Removal of batch effects using distribution-matching residual networks. *arXiv*, 2017.
- [11] Opg batch effect. <https://www.megavisionscan.com/opg/>.
- [12] Clever hans problem. <https://towardsdatascience.com/deep-learning-meet-clever-hans-3576144dc5a9>.
- [13] Adversarial examples and adversarial training. https://www.youtube.com/watch?v=CIfsB_EYsVI&t=3045s, 2017.
- [14] M. S. Baghshah. Deep learning adversarial learning. *Deep Learning Course Slides 2020*, 2020.
- [15] Difference in train and test distribution. <https://www.youtube.com/watch?v=sfk5h0yC67o>, 2017.
- [16] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *CoRR, abs/1706.06083*, 2017.
- [17] X. Yuan, P. He, Q. Zhu, and X. Li. Adversarial examples: Attacks and defenses for deep learning. *arXiv*, 2018.
- [18] J.-F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal. Numerical optimization: theoretical and practical aspects. *Springer Science Business Media*, 2006.
- [19] A. Kurakin, I. J. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *arXiv*, 2017.
- [20] PgD attack. <https://towardsdatascience.com/know-your-enemy-7f7c5038bdf3>.
- [21] PgD image. https://commons.wikimedia.org/wiki/File:Projected_Gradient_Descent.webm.
- [22] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. The limitations of deep learning in adversarial settings. *n 2016 IEEE European Symposium on Security and Privacy (EuroSP)*, 2016.

- [23] W. Brendel, J. Rauber, and M. Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv*, 2017.
- [24] Deep fool attack. <https://towardsdatascience.com/deepfool-a-simple-and-accurate-method-to-fool-deep-neural-networks-17e0d0910ac0>.
- [25] C. Guo, J. Gardner, Y. You, A. G. Wilson, and K. Weinberger. Simple black-box adversarial attacks. In *International Conference on Machine Learning*, 2019.
- [26] Universal adversarial attack. https://www.researchgate.net/figure/A-universal-adversarial-example-fools-the-neural-network-on-images-Left-images-original_fig3_321860123.
- [27] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. *arXiv*, 2017.
- [28] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. Distillation as a defense to adversarial perturbations against deep neural networks. *arXiv*, 2016.
- [29] Defensive distillation image. https://www.researchgate.net/figure/Defensive-Distillation-in-a-nutshell_fig2_348009088.
- [30] Introducing neural structured learning in tensorflow. <https://blog.tensorflow.org/2019/09/introducing-neural-structured-learning.html>.
- [31] M. Andriushchenko and M. Hein. Provably robust boosted decision stumps and trees against adversarial attacks. *Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 2019.
- [32] M. Andriushchenko and M. Hein. Provably robust boosted decision stumps and trees against adversarial attacks. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 2019.
- [33] N. Kumari, M. Singh, A. Sinha, H. Machiraju, B. Krishnamurthy, and V. N. Balasubramanian. Harnessing the vulnerability of latent layers in adversarially

- trained models. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, 2019.*
- [34] M. A. F. Croce and M. Hein. Provable robustness of relu networks via maximization of linear regions. *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16- 18 April 2019, Naha, Okinawa, Japan, 2019.*
- [35] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana. Certified robustness to adversarial examples with differential privacy. *In 2019 IEEE Symposium on Security and Privacy (SP), 2019.*
- [36] P. Benz, C. Zhang, A. Karjauv, and I. S. Kweon. Robustness may be at odds with fairness. *An empirical study on class-wise accuracy, 2020.*
- [37] Y.-Y. Yang, C. Rashtchian, H. Zhang, R. Salakhutdinov, and K. Chaudhuri. A closer look at accuracy vs. robustness. *Advances in Neural Information Processing Systems, 2020.*
- [38] J. Gilmer, N. Ford, N. Carlini, and E. D. Cubuk. Adversarial examples are a natural consequence of test error in noise. *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, 2019.*
- [39] H. Y. Moghaddam. Application of adversarial training in medical signals. 2020.
- [40] X. Li, H. Xiong, X. Li, X. Wu, X. Zhang, J. Liu, J. Bian, and D. Dou. Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. *arXiv, 2021.*
- [41] O. Özdenizci, Y. Wang, T. Koike-Akino, and D. Erdoğmuş. Adversarial deep learning in eeg biometrics. *IEEE Signal Processing Letters, 2019.*
- [42] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *arXiv, 2015.*
- [43] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.*

- [44] U. Shaham. Batch effect removal via batch-free encoding. *bioRxiv*, 2018.
- [45] *Deep Neural Network based Malicious Network Activity Detection Under Adversarial Machine Learning Attacks*, volume 5805 of *LNCS*. Springer, 2020.
- [46] O. Abbasi and J. Gross. Beta-band oscillations play an essential role in motor-auditory interactions. *Human Brain Mapping*, 2020.
- [47] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Ni-jholt, and I. Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 2011.
- [48] Magnetoencephalography. <https://en.wikipedia.org/wiki/Magnetoencephalography>.
- [49] Electroencephalography. <https://en.wikipedia.org/wiki/Electroencephalography>.
- [50] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 2008.
- [51] X. Han, Y. Hu, L. Foschini, L. Chinitz, L. Jankelson, and R. Ranganath. Deep learning models for electrocardiograms are susceptible to adversarial attack. *Nature Medicine*, 2020.
- [52] E. Wong and J. Z. Kolter. Learning perturbation sets for robust machine learning. *CoRR*, 2020.

واژه‌نامه

Maximum Mean.....	بیشینه میانگین اختلاف	الف
Discrepancy	ارزیابی درهم	
بررسی احساسات توسط سیگنال های تنکرده	ارزیابی کنارگذاشتن تکی	
Database For Emotion Analysis Using Phys-	Leave-One-Out.....	
iological Signals	Evaluation	
Arousal	انتقال پذیری	
برانگیختگی	آسیب تخاصمی	
High	آموزگار	
بala	اثر انگشت	
Clip.....	Fingerprint.....	
بهینه ساز	الکتروانسفالوگرافی	
Optimizer	Electroencephalography	
بررسی مولفه های اساسی .	آستانه	
Principal Component Analysis	Threshold	
برچسب زدن	Validation	
Labeling	ارزیابی	
برچسب زدن	استراتژی	
Low	Batch Size	
Convolutional.....	اندازه دسته	
پ	Batch Effect	
پایین	اثر دسته ای	
پیچشی	ب	
ت		
Neurofeedback.....	باخورد عصبی	
Overfitting	بیش برازش	
Cross Entropy.....	بی نظمی تقاطعی	

Student	دانش آموز	Defensive Distillation	تقطیر دفاعی
Binary	دودویی	Normal Distribution	توزیع نرمال
Classification Accuracy	دقت دسته بندی	Laplace Distribution	توزیع لاپلاس
Epoch	دوره	Fully Connected	تماماً متصل

ر

FGSM	روش علامت گرادیان سریع
Basic Iterative Method	روش تکرار شونده ساده
Fine Grained	ریزدانه
Encoder	رمزگذار
Decoder	رمزگشا

س

Policy	سیاست
Deep Neural Networks	شبکه های عمیق
Convolutional Neural	شبکه عصبی پیچشی
	Network

ظ

Plate	ظرف
Valence	ظرفیت

ع

Functional	عملکردی
------------------	---------------

Defensive Distillation	تقطیر دفاعی
Normal Distribution	توزیع نرمال
Laplace Distribution	توزیع لاپلاس
Fully Connected	تماماً متصل

ج

White Box	جعبه سفید
Black Box	جعبه سیاه
Session	جلسه
One Step Targeted	حمله یک گامی هدف دار
Attack	Attack

ح

Black Box Attack	حمله های جعبه سیاه
Boundary Attack	حمله مرز تصمیمی
Adversarial Attack	حمله تخاصمی
Long Short Term	حافظه کوتاه مدت طولانی
Memory	Memory

خ

Piecewise Linear	خطی تکه ای
Manifold	خمیده

د

Classification	دسته بندی
Data Augmentation	داده افزایی

نزول گرادیان افکنده شده Projected Gradient

Descent (PGD)

غ

نقشه برجستگی مبتنی بر ماتریس ژاکوبین . Jacobian Based Saliency Map	غیرخطی Non-Linear
نمونه های تخاصمی Adversarial Examples	غیرفعال Passive
نرخ یادگیری Learning Rate	غلبه Dominance
Music Video..... نماهنگ	
Neuron..... نورون	
Data Normalization نرمال سازی داده	فعال Active

و

وظیفه پایین دستی Downstream Task

ک

کاربردی Functional

ه

هموارسازی تصادفی ... Randomized Smoothing	گام Stride
هويت فرد	گسترش Extension

ی

یادگیری عصبی ساختارمند ... Neural Structured Learning	مسئله اسب باهوش Clever Hans Problem
یادگیری تقویتی Reinforcement Learning	مخفى Hidden
یادگیری ماشین متخصص ... Adversarial Machine Learning	معیار فاصله Distance Measure
یادگیری میانبر Shortcut Learning	میرایی حسی Sensory Attenuation
	مگنتواسفالوگرافی Magnetoencephalography
	مولد Generator

م

یادگیری ساختارمند ... Neural Structured Learning	مسئله اسب باهوش Clever Hans Problem
یادگیری تقویتی Reinforcement Learning	مخفى Hidden
یادگیری ماشین متخصص ... Adversarial Machine Learning	معیار فاصله Distance Measure
یادگیری میانبر Shortcut Learning	میرایی حسی Sensory Attenuation
	مگنتواسفالوگرافی Magnetoencephalography
	مولد Generator

ن

Abstract

Nowadays, Artificial Intelligence has a key factor in a variety of fields such as medicine. beside the incredible progress in the hardware field, the use of deep neural networks increased. Deep neural networks have the ability of identify and estimate data distribution, furthermore they have the ability of extracting high and low level feature that help them to predict highly accurate output.

In medicine the main challenge is the batch effect. the batch effect related to the difference distribution which caused by variety of setting in medical device or different instruction for lab technicians or different situation or etc.

In recent years, The vulnerability of Machine learning models to adversarial attack has been investigated. Adversarial attacks have the ability to modify the model predictions with small perturbation in the data(sometimes in targeted way or sometimes in Stochastic way). Robust learning is the field that deals with adversarial attacks and tries to increase the power of machine learning models.

In this research, By using targeted and non-targeted projected gradient decent attack, we try to effectively increase the data in order to remove batch effect. In this paper, we are going to equalize batch characteristic of data to prevent deep model from learning batch features and force them to learn proper features.

We will use two datasets in the field of brain signals which have batch effects, and we will show that, by using adversarial learning, accuracy of deep model can be increased by several percent in unseen data which may have diffrent distribution against train dataset.

Keywords: Adversarial Learning, Batch Effect, Deep Neural Networks, Robust Learning



Sharif University of Technology
Department of Computer Engineering

B.S. Thesis

Application Of Adversarial Learning In Medical Image

By:

Sina Kazemi

Supervisor:

Dr. MohammadHossein Rohban

July 2022