مروری بر آسیب‌پذیری شبکه‌های عصبی عمیق نسبت‌به نمونه‌های خصمانه و رویکردهای مقابله با آن‌ها

محمد خالوئی، محمدمهدی همایون‌پور* و مریم امیرمزلقانی
دانشکدهٔ مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر، تهران، تهران

چکیده

امروزه شبکه‌های عصبی به‌عنوان بارزترین ابزار مطرح در هوش مصنوعی و یادگیری ماشین شناخته‌شده، و در حوزه‌های مالی و بانک‌داری، کسب‌وکار، تجارت، سلامت، پزشکی، بیمه، رباتیک، هواپیمایی، خودرو، نظامی و سایر حوزه‌ها استفاده می‌شوند. در سال‌های اخیر موارد بی‌شماری از آسیب‌پذیری شبکه‌های عصبی عمیق نسبت‌به حملاتی مطرح شده که به‌طور غالب، با افزودن اختلالات جمع‌شونده و غیرجمع‌شونده بر دادهٔ ورودی ایجاد می‌شوند. این اختلالات با وجود نامحسوس‌بودن در ورودی از دیدگاه عامل انسانی، خروجی شبکهٔ آموزش‌دیده را تغییر می‌دهند. به اقداماتی که شبکه‌های عصبی عمیق را نسبت‌به حملات مقاوم می‌کنند، دفاع اطلاق می‌شود. برخی از روش‌های حمله، مبتنی‌بر ابزارهایی نظیر گرادیان شبکه نسبت‌به ورودی، در پیِ شناسایی اختلال هستند و برخی دیگر به تخمین آن ابزارها می‌پردازند و در تلاشند حتی بدون داشتن اطّلاعاتی از آن‌ها، به اطلاعاتشان دست پیدا کنند. رویکردهای دفاع نیز برخی روی تعریف تابع هزینهٔ بهینه و همچنین، معماری شبکهٔ مناسب، و برخی دیگر بر جلوگیری، یا اصلاح داده قبل از ورود به شبکه متمرکز می‌شوند. همچنین، برخی رویکردها به تحلیل میزان مقاوم‌بودن شبکه نسبت‌به این حملات و ارائهٔ محدودهٔ اطمینان متمرکز شده‌اند. در این مقاله سعی شده‌است جدیدترین پژوهش‌ها در زمینهٔ آسیب‌پذیری شبکه‌های عصبی عمیق بررسی و نقد شوند، و با انجام آزمایش‌هایی نسبت‌به بررسی کارایی هریک، اقدام و آنها را با هم مقایسه شده‌اند. در آزمایش‌های انجام‌شده در بین حملات محصورشده به l_∞ و l_2، به‌ترتیب روش PGD و روش DeepFool کارایی بالاتری دارند. زمان اجرا نیز از نکاتی است که تحلیل شده و نشان داده شده‌است که باتوجه‌به برتری روش‌های PGD و DeepFool، این دو روش برای اجرا، مدت‌زمان بیشتری نسبت‌به سایر روش‌های هم‌ردیف خود نیاز دارند و این میزان در DeepFool از همهٔ روش‌های حمله بیشتر است. همچنین، به مقایسهٔ برخی از رویکردهای پرکاربرد دفاع نسبت‌به نمونه‌های خصمانه نیز پرداخته‌شد؛ که از بین روش‌های مبتنی‌بر نواحی محصورشده به l_∞ حول داده، روش آموزش خصمانه، مبتنی‌بر PGD با شاخص‌های معیّن، از سایر روش‌ها بهتر و در مقابل اغلب روش‌های حمله مقاوم بوده‌است. گفتنی است که روش‌های مختلف حملهٔ خصمانه و همچنین، روش‌های مختلف دفاع نسبت‌به آن حملات که در این مقاله بررسی شده‌است، از طریق نشانیhttps://github.com/khalooei/ars در دسترس علاقه‌مندان قرار دارند.

کلیدواژه‌ها: آسیب‌پذیری شبکه‌های عصبی، مقاوم‌سازی، حمله، دفاع، شبکه‌های عصبی

A Survey on Vulnerability of Deep Neural Networks to Adversarial Examples and Defense Approaches to Deal with Them

Mohammad Khalooei, Mohammad Mehdi Homayounpour*
Maryam Amirmazlaghani
[1]Computer Engineering department, Amirkabir University of Technology, Tehran, Tehran,

Abstract

Nowadays the most commonly used method in various tasks of machine learning and artificial intelligence are neural networks. In spite of their different uses, neural networks and Deep neural networks (DNNs) have some vulnerabilities. A little distortion or adversarial perturbation in the input

* نویسندهٔ عهده‌دار مکاتبات
* Corresponding author

• تاریخ ارسال مقاله: ۱۳۹۹/۱۰/۳۰ • تاریخ پذیرش: ۱۴۰۲/۴/۱۴ • تاریخ انتشار: ۱۴۰۲/۷/۳۰ • نوع مطالعه: کاربردی

[DOI: 10.61186/jsdp.20.2.113]

data for both additive and non-additive cases can be led to change the output of the trained model, and this could be a kind of DNN vulnerability. Despite the imperceptibility of the mentioned disturbance for human beings, DNN is vulnerable to these changes.

Creating and applying any malicious perturbation named "attack", penetrates DNNs and makes them incapable of doing the duty assigned to them. In this paper different attack approaches were categorized based on the signal applied in the attack procedure. Some approaches use the gradient signal for detecting the vulnerability of DNN and try to create a powerful attack. The other ones create a perturbation in a blind situation and change a portion of the input to create a potential malicious perturbation. Adversarial attacks include both black-box and White-box situations. White-box situation focuses on training loss function and the architecture of the model, but black box situation focuses on the approximation of the main model and dealing with the restriction of the input-output model request.

Making a deep neural network resilient against attacks is named "defense". Defense approaches are divided into three categories. One of them tries to modify the input, the other one makes some changes in the developed model and also changes the loss function of the model. In the third defense approach some networks are first used for purification and refinement of the input before passing it to the main network. Furthermore, an analytical approach was presented for the entanglement and disentanglement representation of inputs of the trained model. The gradient is a very powerful signal usually used in learning and an attacking approach. Besides, adversarial training is a well-known approach in changing a loss function method to defend against adversarial attacks.

In this study, a critical literature review has been carried out to summarize and evaluate the latest researches on the vulnerability of DNN. Literature and our experiments indicate that the projected gradient descent (PGD) and DeepFool methods are powerful approaches in the l_2 and l_∞ bounded attacks, respectively. Also, our experiments imply that the PGD and DeepFool are much more time-consuming than the other methods. The DeepFool method is also the most time-consuming approach among all approaches discussed in this paper. In defensive concept, different experiments were conducted to compare different attacks in the adversarial training approaches. Adversarial training is the best defense approach which has been introduced till now, and our experimental results indicate that the PGD is much more effective than fast gradient sign method (FGSM) and Faster FGSM (FFGSM) in adversarial training and cover more generalization of the trained model on the predefined dataset. Also, it is proved that the adversarial training is more time-consuming than pure training. Extended experiment results and more information about this paper are available on https://github.com/khalooei/adversarial_robustess_stack.

Keywords: **vulnerability of neural network, robustness, attack, defense, neural network**

۱- مقدمه

یادگیری ماشین یکی از حوزه‌های پرطرفدار و پرکاربرد در حوزهٔ هوش مصنوعی است. در سال‌های اخیر، از زیرشاخه‌های این حوزه به نام یادگیری عمیق استقبال شده‌است. این شبکه‌ها توانسته‌اند روش‌های شناخته‌شده مانند درخت تصمیم، ماشین بردار پشتیبان و بیشتر رویکردهای پیشین را پشت سرگذاشته و در فضای استخراج ویژگی، دسته‌بندی و وظایف مشابه پیشتاز باشند [۱] و [۲].

اندیشهٔ استفاده از لایه‌های بیشتر در شبکه‌های عصبی و عمیق شدن آن‌ها از الهاماتی است که از علوم اعصاب[1] برگرفته شده‌است [۳]–[۵]. شبکه‌های عصبی با افزایش لایه‌ها، فرآیند یادگیری و درک ویژگی‌های دادهٔ ورودی را بهتر انجام می‌دهند. بالارفتن تعداد لایه‌ها و انواع آنها در شبکه‌های عصبی عمیق موجب بیشترشدن پیچیدگی و بزرگ شدن فضای پارامتری می‌شود. افزایش شاخص‌ها نیز به‌نوبهٔ خود موجب طولانی و سخت‌شدن جستجو در راستای رسیدن به پاسخ سراسری بهینه می‌شود [۱]، [۲] [۵]. طراحی ساختار و ارائهٔ معماری شبکه‌های عصبی عمیق، تعداد لایه‌ها، نحوهٔ چینش آنها و سایر شاخصه‌های شبکه، امری پیچیده و بعضاً مبتنی‌بر تجربه است. همین امر موجب شده تا در کاربردهای حساس، استفاده از این الگوهای یادگیری قابل‌تأمّل شود. درواقع، نباید تنها به دقت، صحّت و سایر معیارهایِ ارزیابی مرسوم شبکه‌های عصبی عمیق توجه کرد؛ و نیاز است از نظر مقاوم‌بودن نسبت‌به حملات خصمانه نیز ارزیابی شوند [۶].

به بیان دیگر، کارایی بالای شبکه‌های عصبی عمیق و از طرفی استفادهٔ فزاینده از آنها، انگیزه‌ای را در افراد شرور و متخاصم ایجاد کرده‌است تا با دستکاری داده‌های ورودی به شبکه‌های عصبی عمیق الگو، عملکرد آنها را مختل و در کار خود (مانند دسته‌بندی) دچار اشتباه کند، به این کار حملهٔ خصمانه گفته‌می‌شود. حملات خصمانه و در نتیجه، تشخیص اشتباه توسط الگو ممکن است عواقب وخیمی را به‌دنبال داشته‌باشد. عدم تشخیص درست تابلوی «ایست»

[1] Neuro science

[DOI: 10.61186/jsdp.20.2.113]

توسط برنامهٔ بینایی رایانه در یک خودروی هوشمند، ایمنی خودرو و سرنشینان آن را به خطر می‌اندازد. اشتباه در تشخیص بدافزار از غیربدافزار می‌تواند امنیت سامانه‌ها و شبکه‌های رایانه‌ای یک سازمان را با مشکل اساسی روبه‌رو کند. فعال‌سازی اشتباه یک سلاح که با فرمان‌های صوتی اداره می‌شود، می‌تواند به نتایج مرگ‌باری منتهی شود. همهٔ اینها مثال‌هایی از اهمیت مقابله با حملات خصمانه و لزوم مقاوم‌سازی شبکه‌های عصبی عمیق نسبت‌به این حملات هستند.

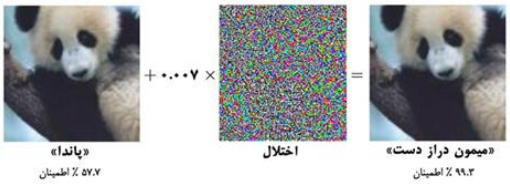
یادگیری ماشین تخاصمی[1] بر مقاوم‌سازی الگوهای یادگیری ماشین نسبت‌به حملات تمرکز دارد. درصد پاسخ درست الگوهای یادگیری ماشین با وجود حملات خصمانه، تعیین‌کنندهٔ میزان مقاومت آن‌ها نسبت‌به این گونه حملات است. متخاصم در تلاش است تا با استفاده از دادهٔ مخرب[2]، الگوهای یادگیری ماشین، به‌ویژه شبکه‌های عصبی را فریب دهد.

مسائلی که مبتنی‌بر یادگیری ماشین هستند، بیشتر دارای فرضیات خاصی هستند. برای مثال، یکسانی شرایط در زمان آموزش و آزمون یکی از مهم‌ترین این فرضیات به شمار می‌رود. درواقع، منظور این است که باید توزیع آماری داده‌های آموزشی با توزیع آماری داده‌های آزمون تا حد بالایی نزدیک باشد. این نزدیکی موجب می‌شود الگوی آموزش‌دیده قابلیت استفاده در محیط آزمون را نیز داشته‌باشد. اما در حملات تخاصمی، در عمل، متخاصم با زیرکی ویژه‌ای شرایط محیط آزمایش را در قیاس با شرایط محیط آموزش تغییر می‌دهد و همین امر موجب به اشتباه افتادن دسته‌بند و بروز مشکلات عدیده‌ای می‌شود [۷]، [۸] و [۱۸].

یکسانی محیط‌های آموزش و آزمون، از مواردی است که منجر می‌شود، متخاصم این امکان را داشته‌باشد با اعمال تغییرات نامحسوس از دید عامل انسانی در دادهٔ ورودی، دسته‌بند را دچار خطا کند. البته به‌الزام تفاوت توزیع دادگان آموزش و آزمون تنها دلیل آن نیست و دلایل مختلف دیگری از قبیل معماری نامناسب شبکه، تابع هزینه و مانند آن نیز می‌توانند از دیگر دلایل بروز این مسئله باشند. پژوهش‌های علمی بسیاری به بررسی انواع حمله و روش‌های دفاع و جوانب مختلف آن پرداخته‌اند [۶]، [۸]–[۱۲]. بخش دوم متمرکز بر مفهوم حمله، به شبکه‌های عصبی و مرور رویکردهای مختلف آن پرداخته‌است. در بخش سوم، به تعریف دفاع و بیان روش‌های مختلف دفاع از حملات خصمانه اشاره شده، در بخش چهارم به بحث و بررسی روش‌های مختلف، و در بخش پنجم نیز به جمع‌بندی اشاره شده‌است.

۲- تعریف حمله[3]

به تغییرات بسیار اندک که در عین نامحسوس‌بودن از دیدگاه عامل انسانی، موجب اشتباه در تشخیص ردهٔ دادهٔ ورودی توسط الگوی یادگیری ماشینی می‌شود، اختلال[4]، و به اقداماتی که برای افزودن اختلال و به‌اشتباه‌انداختن الگو صورت می‌گیرد، حمله گفته‌می‌شود. حملات موجب فریب الگوی یادگیری ماشین می‌شود و کلاس خروجی را به‌ازای دادهٔ ورودی تغییر می‌دهند. گفتنی است که با استفاده از حملات بی‌شمار به الگوهای یادگیری ماشین، آسیب‌پذیری[5] این الگوها به روش‌های مختلفی قابل‌ارزیابی است [۶]، [۸]، [۱۱] و [۱۸].



(شکل-۱) یک حملهٔ نمونه با الگوی شبکهٔ عصبی عمیق [۸]

(Figure 1) A sample attack on the Deep Neural Network model [8]

برای مثال، در دسته‌بندی داده‌ها، اگر دادهٔ ورودی با اندکی تغییرات نامحسوس تلفیق شود، الگوی مسئله، فریب می‌خورد و کلاس خروجی تغییر می‌کند. مثالی از دسته‌بندی تصاویر را در شکل (۱) در نظر بگیرید. در این شکل، تصویر ورودی $\boldsymbol{x}$ با دقت ۵۷/۷ درصد «پاندا» تشخیص داده‌می‌شود. افزودن تغییرات اندک (نامحسوس از نظر عامل انسانی)، موجب شده‌است که این داده با دقت اطمینان ۹۹/۳ درصد «گونه‌ای میمون» تشخیص داده شود. [۸].

این حملات متناسب با شرایط و اهداف می‌توانند متفاوت باشند. در ادامه به جزئیات اهداف و انواع آن‌ها پرداخته شده‌است.

اغلب پیداکردن یک اختلال بهینه، نیازمند حل مسئلهٔ بهینه‌سازی است؛ ولی در برخی مواقع متناسب با شرایط، ممکن است از روش‌های سعی و خطا نیز استفاده شود. اغلب یک اختلال قادر به تغییر ردهٔ یک داده خاص است و قابلیت عام‌بودن ندارد. اما می‌توان با روش‌هایی

[1] Adversarial Machine Learning
[2] Malicious
[3] Attack
[4] Perturbation
[5] Vulnerability

[DOI: 10.61186/jsdp.20.2.113]

اختلالی عام‌منظوره‌ای[1] (همگانی) را تولید کرد که پس از اعمال روی مجموعه‌ای از داده‌ها موجب فریب الگو شود. صورت کلی مسئله حمله به‌صورت رابطهٔ (۱) تبیین می‌شود:

$$\min_{\eta} \|\eta\|_p \quad s.t.\, F(x+\eta) \neq F(x) \qquad (۱)$$

در رابطهٔ (۱)، منظور از x دادهٔ ورودی، η همان اختلال و F دسته‌بند یا الگوی فراگرفته‌شدهٔ مسئله است. منظور از $\|\cdot\|_p$ نیز نرم l_p است. حملات با استفاده از نظرات مختلف، اختلال‌ها را پیدا می‌کنند. در بخش‌های آینده انواع روی‌کردهای گوناگون برای یافتن اختلال بیان شده‌است [۷] و [۸].

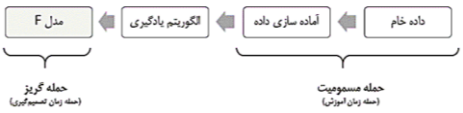
انواع حملات

دیدگاه‌های گوناگونی برای تقسیم‌بندی روش‌های مختلف حمله بیان شده‌است. در این پژوهش سعی شده در بهینه‌ترین حالت این تقسیم‌بندی ارائه شود. این تقسیم‌بندی از سه دیدگاه مرحله و زمان حمله، اطلاعات در دسترس متخاصم و اهداف کلی متخاصم انجام شده‌است.

از دیدگاه مرحله وقوع حمله

یکی از موردتوجه‌ترین نکات در حمله، مرحلهٔ وقوع حمله است. حمله می‌تواند به‌صورت مستقیم روی الگو تأثیر داشته‌باشد. زمانی‌که حمله روی الگوی آموزش‌دیده انجام شود (مرحلهٔ آزمون)، به آن حملهٔ گریز[2] گفته می‌شود و یکی از مرسوم‌ترین حملات است. و زمانی‌که حمله با ایجاد تغییرات مخرب بر روی دادهٔ آموزشی (مرحله آموزش) صورت گیرد، به آن حملهٔ مسمومیت[3] گفته‌می‌شود. همان‌طورکه در (شکل-۲ مشاهده می‌شود، حملات گریز، اغلب، در مرحلهٔ آزمون (روی الگوی F) انجام می‌شوند. حملات مسمومیت برخلاف حملات گریز، در زمان آموزش و یا آماده‌سازی داده انجام می‌شوند.

از دیدگاه اینکه متخاصم چه میزان داده و اطلاعات از سیستم مورد هدف دراختیار دارد، حملات به دو دسته جعبه‌سفید[4] و جعبه‌سیاه[5] تقسیم‌بندی می‌شوند. در ادامه به بیان تفصیلی این دو دسته حمله خواهیم پرداخت [۶] و [۱۱].



(شکل-۲) انواع حملات از دیدگاه مرحله و زمان وقوع حمله [۶]، [۱۲]

(Figure 2) Types of attacks from the point of view of the stage and time of the attack [6], [12]

از دیدگاه اطلاعات دردسترس

حملهٔ جعبه‌سفید

در حملهٔ جعبه‌سفید فرض بر این است که متخاصم اطلاعات دقیقی از الگو در اختیار دارد. اگر این اطلاعات تنها به اطلاعات الگوی نهایی ختم شود، این حمله از نوع حملهٔ گریز یا حملهٔ زمان تصمیم‌گیری است. اگر این اطلاعات تنها به جزئیات یادگیری نظیر دادهٔ آموزشی و موارد مشابه مربوط باشد، حمله از نوع حملهٔ مسمومیت است. در حملهٔ مسمومیت، با روش‌های گوناگونی نظیر تزریق دادهٔ آلوده به دادگان آموزشی تمیز، تغییر برچسب برخی از دادگان آموزشی و به‌طورکلی مختل‌کردن روال آموزش، می‌توان با تغییر در الگو بر خروجی نهایی تأثیر مخرّب داشت. دسترسی متخاصم به اطلاعات کامل از الگو به‌معنی دسترسی او به اطلاعات دقیق شاخص‌های الگوی نظیر معماری، ویژگی‌های فراگرفته‌شده (وزن‌ها)، بهینه‌ساز و شاخص‌های الگوست [۶] و [۱۱].

حملهٔ جعبه‌سیاه

برخلاف حملهٔ جعبه‌سفید، در این نوع حمله، متخاصم هیچ‌گونه اطلاع دقیقی از الگو یا فرآیندهای آموزشی نظیر آماده‌سازی داده و جزئیات آن در اختیار ندارد. در ساده‌ترین سطح، متخاصم می‌تواند به‌صورت محدود به الگو ورودی دهد و خروجی آن را دریافت کند. برای مثال، اگر ورودی x به‌عنوان یک درخواست به سامانه داده شود، الگو بر اساسِ ورودی درخواست‌شده، خروجی‌های متناظر را تولید می‌کند ($y_{pred} = F(x)$). متخاصم دراین‌صورت می‌تواند بر پایهٔ y_{pred} دریافتی، تخمین‌هایی از F داشته‌باشد. در سطح بعد ممکن است متخاصم اطلاعاتی نظیر رده‌های فراگرفته‌شده در الگو، ویژگی‌ها و شاید اندکی از جزئیات یادگیری را در اختیار داشته‌باشد؛ اما اطلاعاتی از دادهٔ آموزشی نداشته‌باشد.

درخصوص محدودیّت دسترسی به دادگان آموزشی ممکن است متخاصم هیچ‌گونه اطلاعاتی از الگو در اختیار نداشته‌باشد، ولی به تعدادی نمونهٔ آموزشی که الگو با آن آموزش دیده، دسترسی داشته‌باشد. در نتیجه، زمانی‌که

[1] Universal
[2] Evasion
[3] Poisoning
[4] White-box
[5] Black-box

[DOI: 10.61186/jsdp.20.2.113]

متخاصم اطّلاعات تقریبی یا دقیق از دادگان آموزشی داشته‌باشد، حمله می‌تواند به نوعی از حملهٔ جعبه‌سفید تبدیل شود؛ زیرا دراین‌صورت، متخاصم قادر خواهدبود با استفاده از دادگان آموزشیِ در دسترس، الگویی مشابه الگوی اصلی آموزش دهد [۶] و [۱۱].

اغلب بیان می‌شود که حملات جعبهٔ سیاه به اصل انتقال‌پذیری متکی هستند [۱۳]. این بدان معنی است که می‌توان با استفاده از نمونهٔ خصمانهٔ یک الگو، الگوی دیگری را هدف قرار داد. درواقع، باتوجه‌به محدودیت دسترسی به الگو در حملات جعبه‌سیاه، اغلب تلاش می‌شود تا الگویی جایگزین[1] به گونه‌ای آموزش ببینند که دقت و میزان پاسخ‌گویی آن به الگوی اصلی نزدیک باشد؛ سپس، با حملهٔ جعبه‌سفید به الگوی جایگزین، نمونه‌های خصمانه تولید، و پس از آن، با استفاده از اصل انتقال‌پذیری، نمونه‌های خصمانه تولیدی روی الگوی هدف، آزمایش و به‌عبارت دیگر، به الگوی هدف حمله می‌شود.

از دیدگاه هدف

حمله به الگو موجب تغییر پاسخ خروجی و تشخیص اشتباه در تصمیم‌گیر، یا کاهش سطح اطمینان پاسخ می‌شود. باتوجه‌به اهداف حمله، دسته‌بندی گوناگونی از حملات، قابل ارائه است. اما به‌طورکلی می‌توان اهداف مختلف یک متخاصم را در دو دستهٔ حملات هدف‌مند[2] و حملات کاهش سطح اطمینان[3] خلاصه کرد.

حملات هدف‌مند

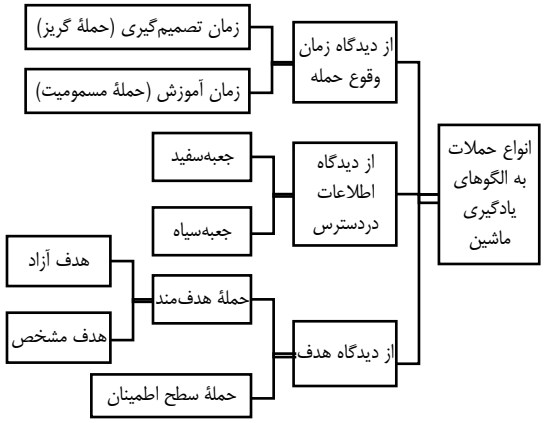
در حملات هدفمند، هدف فریب‌دهنده این است که پاسخ الگو را از یک رده به ردهٔ دیگر تغییر دهد. این حمله خود، می‌تواند به دو دستهٔ آزاد و مشخص تقسیم‌بندی شود. در حملهٔ هدف آزاد، مقصود این است که ردهٔ دادهٔ ورودی به غیر از ردهٔ اصلی خودش تغییر یابد. در هدف مشخص، مقصود این است که ردهٔ دادهٔ ورودی به ردهٔ خاصّی تغییر یابد. همچنین، حالت خاص‌تری وجود دارد و آن این است که ردهٔ مبدأ مشخصی به ردهٔ مقصد مشخص دیگری تغییر یابد. در برخی پژوهش‌ها نظیر [۳۴] و [۳۶]–[۳۸]، هدف واضح و مشخص است؛ ولی در برخی دیگر نظیر [۱۷]–[۱۹] در یک گام، نزدیک‌ترین هدف تولید، سپس، حمله انجام می‌شود.

حملات سطح‌اطمینان

در حملات سطح اطمینان برخلاف حملات هدف‌مند، هدف آن است که با افزایش خطای پیش‌بینی الگو، سطح اطمینان پاسخ آن کاهش یابد و میزان خطای خروجی الگو نسبت‌به خروجی سابق تا حد امکان افزایش یابد. برای مثال، اگر سطح اطمینان دسته‌بند، پیشتر برای داده‌های ورودی از ردهٔ مدنظر ۹۸ درصد بوده‌باشد، این حمله در تلاش است تا حد ممکن سطح اطمینان آن را کاهش دهد و به‌طور مثال، به زیر ۵۰ درصد برساند. در شکل (۳) تقسیم‌بندی کلی انواع حمله به الگوهای یادگیری ماشین ارائه شده‌است.

تولید نمونهٔ خصمانه[4]

به‌منظور پیداکردن و شناسایی نمونهٔ خصمانه از الگوریتم‌ها و روش‌های متفاوتی استفاده می‌شود. در حالت کلی نیاز است متناسب با نوع حمله، اختلال را تولید، سپس، با دادهٔ اولیه ترکیب کرد. در برخی روش‌ها دو عمل تولید اختلال و ترکیب، با هم انجام می‌پذیرد.



(شکل- ۳) تقسیم‌بندی انواع حمله به الگوهای یادگیری ماشین [۶]، [۱۱] و [۲۱]

(Figure 3) Different types of attack to machine learning models [6], [11], [21]

در برخی روش‌ها نظیر [۲۱]، فرآیند ترکیب اختلال با داده نیز به‌صورت خودکار انجام شده‌است. رابطهٔ تولید نمونهٔ خصمانه به‌صورت رابطهٔ (۲) بیان می‌شود:

$$\hat{x} = x + \underset{\eta}{\operatorname{argmin}}\{\|\eta\| : F(x+\eta) \neq F(x)\} \qquad (۲)$$

در رابطهٔ (۲)، تولید نمونهٔ خصمانه از طریق حل مسئلهٔ بهینه‌سازی انجام می‌شود. یافتن کمینهٔ یک تابع هزینه دارای رویکردهای متنوعی است. در توابع محدّب به‌راحتی می‌توان به کمینهٔ محلی رسید و ازآنجاکه کمینه‌های محلی در توابع محدّب، کمینه‌های سراسری نیز هستند، به‌نسبت و به‌راحتی می‌توان جواب را یافت. اما نکتهٔ اصلی اینجاست که تابع هزینهٔ شبکه‌های عصبی و به‌ویژه، شبکه‌های عصبی عمیق، به‌طور عموم، از نوع

[1] Surrogate
[2] Targeted attack
[3] Confidence reduction attack
[4] Adversarial example

[DOI: 10.61186/jsdp.20.2.113]

غیرخطی و غیرمحدب هستند. این موضوع موجب می‌شود پیداکردن کمینه به‌راحتی صورت نپذیرد. همچنین، بدیهی است که کمینه‌های محلی در توابع غیرمحدب، به‌الزام کمینه‌های سراسری نیستند و همین امر موجب سخت‌شدن و عدم تعریف فرم بسته برای این نوع از مسئله‌ها می‌شود [۷] و [۸].

روش‌های مختلف تولید نمونهٔ خصمانه

حل مسئلهٔ بهینه‌سازی رابطهٔ (۲)، به روش‌های متعددی قابل انجام است. در ادامه سعی شده‌است تا به‌صورت مختصر مروری بر برخی پژوهش‌های مرتبط و کلیدی صورت پذیرد.

روش L-BFGS[1]: *سژدی*[2] و همکاران در پژوهش [۹] پیداکردن اختلال بهینه برای تغییر ردهٔ ورودی به رده‌ای غیر از ردهٔ اصلی را با رابطهٔ (۳) تعریف کردند:

$$\begin{aligned} &\boldsymbol{minimum}\ \|\boldsymbol{\eta}\|_2 \\ &\boldsymbol{s.t}\ \ \boldsymbol{F(x+\eta) \neq F(x)} \\ &\boldsymbol{x+\eta \in [0,1]^m} \end{aligned} \qquad (۳)$$

رابطهٔ بالا را می‌توان به‌صورت رابطهٔ (۴) نیز بازنویسی کرد:

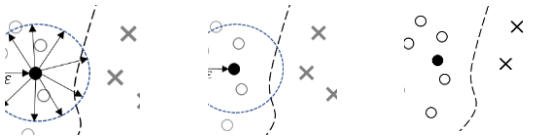
$$\begin{aligned} &\boldsymbol{minimum}\ \boldsymbol{c\|\eta\|_2 + loss_{F,t}(x+\eta)} \\ &\boldsymbol{s.t}\ \ \boldsymbol{x+\eta \in [0,1]^m} \end{aligned} \qquad (۴)$$

در رابطهٔ (۴) منظور از ثابت $\boldsymbol{c}$ ضریب ترکیب خطی است. همچنین، منظور از $\boldsymbol{loss_{F,t}}$ تابع هزینهٔ دسته‌بند $\boldsymbol{F}$ مدنظر به‌صورت هدف‌مند با هدف مشخص $\boldsymbol{t}$ است. آن‌ها در پژوهش خود به‌منظور حل این مسئله از الگوریتم بهینه‌سازی **L-BFGS** [۲۲][3] استفاده کردند. آن‌ها همچنین، به حالت‌های خاصی از آموزش خصمانه[4]، همچون افزودن نمونه‌های خصمانه به مجموعه داده و سپس، آموزش الگو نیز اشاره کردند. گفتنی است که روش پیداکردن نمونهٔ خصمانه آن‌ها دارای هزینهٔ محاسباتی زیادی است.

روش علامت گرادیان سریع (FGSM[5]**):** *گودفلو*[6] و همکاران در پژوهش [۸]، راه حل بهتری برای حل مسئلهٔ پیداکردن نمونهٔ خصمانه معرفی کردند. این رویکرد مبتنی‌بر اعمال تابع علامت روی گرادیان تابع هزینه نسبت‌به ورودی $\boldsymbol{x}$ است. نمونهٔ خصمانه $\boldsymbol{\hat{x}}$ پژوهش آن‌ها توسط رابطهٔ (۵) تولید می‌شود:

$$\boldsymbol{\hat{x} = x + \varepsilon.sign(\nabla_x J(\theta, x, y_{true}))} \qquad (۵)$$

در عبارت بالا، منظور از $\boldsymbol{sign}$ تابع علامت، $\boldsymbol{J}$ تابع هزینه تعریف‌شده برای الگوی یادگیری، $\boldsymbol{y_{true}}$ برچسب متناسب با دادهٔ ورودی $\boldsymbol{x}$ و $\boldsymbol{\varepsilon}$ میزان تأثیر اختلال روی دادهٔ ورودی $\boldsymbol{x}$ در روش **FGSM** است. تعبیر هندسی مسئلهٔ بالا در شکل (۴) به‌صورت گام‌به‌گام از چپ به راست آورده شده‌است:



الف ب ج

(شکل- ۴) تعبیر رویکرد FGSM با اعمال تغییرات به میزان ε به‌صورت مرحله به مرحله تا رسیدن به نمونهٔ خصمانه مطلوب

(Figure 4) An interpretation of FGSM approach by applying ε-value changes step by step until reaching the desired adversarial example

طبق شکل (۴)، مبتنی‌بر نمونه‌ها، در قسمت الف یک نمونه انتخاب می‌شود، سپس، مبتنی‌بر نمونهٔ انتخاب‌شده، این روش قصد دارد در محدودهٔ $\boldsymbol{\varepsilon}$ به‌دنبال تولید اختلالی باشد که اگر به داده اضافه شود، (**شکل**)- ب موجب شود، نمونهٔ تولیدی از مرز بین رده‌ای عبور کند (شکل (۴)- پ)؛ و آن نمونه در محدوده $\boldsymbol{\varepsilon}$ به‌عنوان نمونهٔ خصمانه اعلام شود.

در نگاه آنان دلیل اصلی وجود نمونهٔ خصمانه، فرض امکان جداسازی کلاسه‌ها به‌صورت خطی در فضای با ابعاد بالا بوده‌است. تعریف مسئلهٔ آن‌ها با نرم بی‌نهایت (استفاده از تابع علامت) بیان شده‌است. میاتو و همکاران تعریف مسئله را با استفاده از طبیعی‌سازی نرم $\boldsymbol{l_2}$ به فرم رابطهٔ (۶) بیان کردند [۲۳]:

$$\boldsymbol{\eta = \varepsilon \frac{\nabla J(\theta, x, y)}{\|\nabla J(\theta, x, y)\|_2}} \qquad (۶)$$

در رابطهٔ (۶) سعی شده‌است تا گرادیان با نسبت نرم $\boldsymbol{l_2}$ طبیعی شود. حال این نرم اگر نرم بی‌نهایت باشد، به همان تعریف **FGSM** نسخهٔ اولیه $\boldsymbol{\eta = \varepsilon\, sign(\nabla J(\theta, x, y))}$ تبدیل می‌شود.

روش علامت گرادیان سریع هدف‌مند[7]**:** *کوراکین*[8] و همکاران [۱۱]، قابلیت هدف‌مندشدن نمونهٔ خصمانه را با کمی تغییر در مسئلهٔ بهینه‌سازی **FGSM** ارائه کردند. منظور از هدف‌مندبودن آن است که فریب صورت‌گرفته به‌گونه‌ای باشد تا برچسب رده به برچسب

[1] Limited-Broyden Fletcher Goldfarb Shannon
[2] Szegedy
[3] BFGS نسخهٔ بهینه‌شده و محدودشده حافظه‌ای روش بهینه‌سازی [۷۳] می‌باشد.
[4] Adversarial training
[5] Fast Gradient Sign Method (FGSM)
[6] Goodfellow
[7] Targeted Fast Gradient Sign Method
[8] Kurakin

[DOI: 10.61186/jsdp.20.2.113]

هدف موردنظر تغییر یابد. رابطهٔ ریاضی بهینه‌سازی به‌صورت رابطهٔ (۷) تعریف می‌شود:

$$\hat{x} = x - \varepsilon . sign\left(\nabla_x J(\theta, x, y_{target})\right) \quad (۷)$$

رابطهٔ (۷) همانند رابطهٔ (۵) است، با این تفاوت که در زمان محاسبهٔ گرادیان، تغییرات برچسب هدف مسئله (y_{target}) نسبت‌به تغییرات ورودی x محاسبه می‌شود. همچنین، پس از اعمال تابع علامت، این‌بار از x کم می‌شود. درواقع، برای این است تا بیشینهٔ احتمال خروجی الگو برای ردهٔ y_{target} نسبت‌به ورودی نمونهٔ خصمانه افزایش یابد.

روش‌های تکرارشوندهٔ مبتنی‌بر گرادیان:

کوراکین و همکاران در [۱۱] روش علامت گرادیان سریع تکرارشونده [۱] را معرفی کردند. این پژوهش رویکرد **FGSM** را توسعه داده و در تلاش است تا در تکرارهای بیشتر به‌صورت مرحله‌ای گام بردارد. در نتیجه، سعی می‌کند از این روش به مرز دسته‌بند نزدیک‌تر شده و نمونهٔ بهتر از **FGSM** را برای فریب تولید کند. به همین علت، به آن **iFGSM**[۲] یا **BIM**[۳] گفته‌می‌شود. رابطهٔ ریاضی بهینه‌سازی این پژوهش به‌صورت رابطهٔ (۸) بیان می‌شود:

$$\hat{x}_0 = x \quad (۸)$$
$$\hat{x}_{n+1} = Clip_{x,e}\left\{\hat{x}_n + \alpha\ sign\left(\nabla_x J(\theta, \hat{x}_n, y_{true})\right)\right\}$$

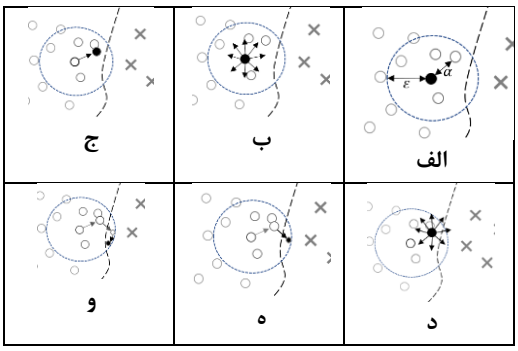
در رابطهٔ (۸) برای‌این‌که دست‌یابی به پاسخ نهایی در چندین گام صورت می‌پذیرد، اولین گام با مقداردهی نمونهٔ اصلی ($\hat{x}_0 = x$) شروع می‌شود. در گام‌های بعدی بر مبنای آخرین نمونهٔ مرحلهٔ قبل ($\hat{x}_n$)، نمونهٔ جدید ($\hat{x}_{n+1}$) تولید می‌شود که در هر گام، **FGSM** را مبتنی‌بر آخرین نمونهٔ مرحلهٔ قبل اجرا نماید. سپس، درصورتی‌که محدودهٔ نمونهٔ خصمانه تولیدی از محدودهٔ تعریف‌شده[۴] عبور کرده‌باشد، مقادیر توسط عملیات $Clip_{x,e}$ به محدودهٔ موردانتظار برمی‌گردند[۵]. این مراحل با تعداد گام مشخص یا تا زمان رسیدن به نمونهٔ خصمانهٔ بهتر ادامه می‌یابد.

در رابطهٔ (۸) منظور از α اندازه گام حرکت و $Clip_{x,e}\{\cdot\}$ عملیات محدودکردن ورودی به محدودهٔ مشخص‌شده است. گفتنی است که هریک از اجزای نمونهٔ تولیدشده، توسط عملیات $Clip_{x,e}\{\cdot\}$، به محدودهٔ $[x_{i,j} - \varepsilon, x_{i,j} + \varepsilon]$ بازگردانی می‌شود. در (شکل- ۴) مراحل گام‌به‌گام الگوریتم **iFGSM** آورده شده‌است.

قسمت الف (شکل- ۴)، نشان‌دهندهٔ نسبت اندازهٔ متغیرهای α و ε در فضای ویژگی موردبحث است. در بخش‌های ب تا ج، روال پیشرفت الگوریتم تکرارشونده نشان داده شده‌است. این روش به نسبت رویکرد **FGSM**، نمونهٔ خصمانهٔ بهتری تولید می‌کند. کوراکین و همکاران در پژوهش خود در خصوص مقاوم‌سازی نیز بحث و به روش آموزش خصمانهٔ پژوهش سژدی و همکاران [۹] نیز استناد کردند. نوع دیگری از رویکرد تکرارشونده **iFGSM**، روش تکرارشوندهٔ کم‌احتمال‌ترین رده[۶] است. کوراکین و همکاران [۱۱] آن را با نام **ILLCM**[۷] ارائه کردند که در رابطهٔ Error! Reference source not found. آورده شده‌است. در این رابطه، رده یا کمترین احتمال با برچسب y_{LL} مشخص شده‌است.

(۹)

$$\hat{x}_0 = x$$
$$\hat{x}_{n+1} = Clip_{x,e}\left\{\hat{x}_n - \alpha sign\left(\nabla_x J(\theta, \hat{x}_n, y_{LL})\right)\right\}$$



(شکل- ۴) گام‌های طرح‌وارهٔ الگوریتم iFGSM به‌صورت مرحله‌به‌مرحله. شکل (الف) نشان‌دهندهٔ نسبت اندازهٔ متغیرهای α و ε در فضای ویژگی و شکل‌های «ه» تا «و» روال پیشرفت الگوریتم تکرارشونده را تا رسیدن به نمونهٔ خصمانهٔ مطلوب نشان می‌دهد.

(Figure 5) Schematic steps of the iFGSM algorithm. Figure A in the feature shows the size ratios of the variables α and ε space and the other steps (figures) show the iterative progress until reaching the desired adversarial example

مَدری[۸] و همکاران در [۲۴] مسئلهٔ پیداکردن اختلال[۹] را با استفاده از روش کمینه گرادیان تصویرشده[۱۰]

[1] Iterative Fast Gradient Sign Method
[2] Iterative FGSM (iFGSM)
[3] Basic Iterative Method (BIM)
[۴] محدودهٔ طبیعی مقادیر ماتریس عددی هر نمونه، به‌عنوان مثال، محدودهٔ طبیعی صفر و یک یا محدوده غیرطبیعی دادهٔ ورودی همانند محدودهٔ [۰,۲۵۵] در تصویر است.
[۵] همهٔ عناصر ماتریس داده و یا همهٔ پیکسل‌ها در داده‌های تصویری به محدودهٔ مقادیر مشخص‌شده محدود و منتقل می‌شوند.

[6] Least-likely (LL) class method
[7] Iterative Least-Likely Class method
[8] Madry
[۹] مسئله پیداکردن اختلال بهینه، یک مسئله بهینه‌سازی دارای قید است.
[10] Projected Gradient Descent

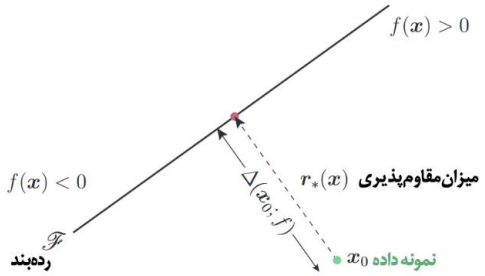
[DOI: 10.61186/jsdp.20.2.113]

به‌جای گرادیان کمینه حل کردند. در روش پیشنهادی آن‌ها، مراحل پیشروی همانند گرادیان سریع تکرارشونده انجام می‌شود و تفاوت زیادی در طول روند پیشروی الگوریتم کلی وجود ندارد.

دستکاری تخاصمی بازنمایی‌عمیق[1]: *صبور* و همکاران در [۲۵]، روش حملهٔ هدفمندی را ارائه کردند که در آن به‌جای دخیل‌کردن فاصلهٔ خروجی حاصل از الگوی (برچسب)، از فاصلهٔ خروجی حاصل از لایه‌های میانی استفاده کردند. درواقع، هدف آن‌ها با ارائهٔ این نظر این بود که بیان کنند نباید اختلافی بین ویژگی‌های استخراج‌شدهٔ لایه میانی دادهٔ تمیز و دادهٔ خصمانه ایجاد شود. رابطهٔ مسئلهٔ بهینه‌سازی آن‌ها در رابطهٔ (۹) آورده شده‌است:

$$\min_{\hat{x}} \ \|\phi_k(x) - \phi_k(\hat{x})\| \quad s.t. \ \ \|x - \hat{x}\|_\infty < \delta \tag{۹}$$

در رابطهٔ (۹) منظور از $\boldsymbol{\phi_k}$، نگاشتی از ورودی به خروجی لایه $\mathbf{k}$ ام شبکه است. همچنین، همانند بیان‌های مختلف دیگر، از مسئلهٔ پیداکردن کمترین اختلال، نرم بی‌نهایت را برای محدودکردن اختلال‌های تولیدی توصیه کردند. گفتنی است که آن‌ها همانند سژدی و همکاران [۹] برای حل مسئله از الگوریتم **LBFGS** استفاده کردند. آن‌ها همچنین، بیان کردند که فرضیات خطی ارائه‌شده در پژوهش گودفلو و همکاران [۸] چندان قابل‌استناد نبوده و به‌ویژه در تغییر خصمانه بازنمایی، کارایی چندانی ندارد و معتقدند این مشکل به ماهیت و معماری شبکه مرتبط است.



(شکل- ۵) نمایی از تصویرکردن نمونهٔ داده در روش DeepFool [۱۵]

(Figure 6) Representation of sample projection in DeepFool method [15]

روش DeepFool: *موسوی دزفولی* و همکاران در [۱۵] روشی جهت حمله با هدف آزاد ارائه کردند. مطابق (شکل- ۵) این روش به‌خاطر استفاده از نرم $\boldsymbol{l_2}$، پیداکردن اختلال را بهتر از روش‌های پیشین نظیر روش **LBFGS** انجام می‌دهد (در روش **LBFGS** تنها به استفاده از نرم $\boldsymbol{l_2}$ در محدودکردن تولید اختلال تأکید دارد). دیدگاه روش آن‌ها با فرض تقریب خطی شبکهٔ عصبی پیش می‌رود که در (شکل- ۵) با $\boldsymbol{f}$ نشان داده شده‌است. در نگاهِ آن‌ها تفکیک یک رده از ردهٔ دیگر توسط ابرصفحه انجام می‌شود. درواقع، آن‌ها با تصویرکردن[2] نمونه به دسته‌بند سعی می‌کنند تا میزان مقاومت‌پذیری را به‌دست‌آورند و از آن برای تولید اختلال استفاده کنند. (شکل- ۵) نمایی از تصویرکردن نمونهٔ داده بر دسته‌بند و گام‌های نخستین روش **DeepFool** را نشان می‌دهد.

به‌منظور تصویرکردن یک نقطه روی دسته‌بند، از رابطهٔ (۱۰) استفاده می‌شود:

$$\boldsymbol{\eta \sim r_* := argmin \|r\|_2 \quad s.t. \ \ sign\big(f(x_0 + r)\big) \neq sign\big(f(x_0)\big) = -\frac{f(x_0)}{\|\nabla f(x_0)\|_2^2}\nabla f(x_0)} \tag{۱۰}$$

در رابطهٔ (۱۰) منظور از $\boldsymbol{r}$ میزان مقاوم‌بودن الگو و به عبارت دیگر، $\boldsymbol{r}$ تعداد قدم‌های موردنیاز جهت رسیدن به مرز دسته‌بند است. با فرضیه‌های بیان‌شده از مسئله، آن‌ها راه‌حل تحلیلی جهت ساده‌سازی مسئله و تولید نمونه‌های خصمانه ارائه کردند. گفتنی است که روش آن‌ها نیز جزء روش‌های تکرارشونده محسوب می‌شود و در گام اول از نمونهٔ تمیز شروع می‌کند. در گام‌های بعدی متناسب با تقریب خطی دسته‌بند، سعی می‌کند روی آن تصویر[3] شود. همین روال تا رسیدن به نمونهٔ خصمانه نیز تکرار می‌شود. با جمع همهٔ $\boldsymbol{r_i}$ ها، می‌توان به اندازهٔ اختلال به دسته‌بند نزدیک شد.

گفتنی است که میزان اختلال نهایی متناسب با میزان مقاومت‌پذیری توسط رابطهٔ (۱۱) به‌دست‌می‌آید. در این رابطه منظور از $\boldsymbol{\varepsilon}$ میزان اختلال موردنیاز جهت عبور از دسته‌بند است که اغلب مقدار آن $\boldsymbol{\varepsilon \ll 1}$ است.

$$\eta = (1 + \varepsilon) \times \hat{r} \tag{۱۱}$$

تعمیم روش آن‌ها در حالت چندرده‌ای نیز بدین صورت است که نخست نزدیک‌ترین مرز دسته‌بند از بین مرز رده‌ها شناسایی و سپس، روی آن مرز دسته‌بند همانند روش دوره‌ای تصویرسازی و ادامهٔ روال انجام می‌پذیرد.

روش نقشه برجستگی مبتنی‌بر ژاکوبین[4]: اغلب روش‌های تولید نمونهٔ خصمانه برای تولید اختلال‌های نامحسوس، با نرم‌های $\boldsymbol{l_2}$ یا $\boldsymbol{l_\infty}$ محدود می‌شوند. *پپرنو*[5] و همکاران در پژوهش [۲۶]، اختلال را به نرم $\boldsymbol{l_0}$ محدود کردند. آن‌ها به‌جای تغییر کل مقادیر (پیکسل‌ها)، به‌دنبال

[1] Adversarial manipulation of deep representations (Rep. Adversary)

[2] Project
[3] Projection
[4] Jacobean Based Saliency Map (JBSM)
[5] Pepernot

[DOI: 10.61186/jsdp.20.2.113]

کمترین تغییر مقدار (تغییر پیکسل) در دادهٔ (تصویر) ورودی بودند. در روش آن‌ها به محاسبهٔ ژاکوبین **F** پرداخته شده‌است. آن‌ها با محاسبهٔ گرادیان خروجی نسبت‌به ورودی در تلاش هستند تا اجزای تأثیرگذار و نقاط برجسته و بارز در دادهٔ ورودی را شناسایی کنند؛ سپس، اختلال خصمانه را مطابق با نقاط برجسته تأثیرگذار روی داده اعمال می‌کنند.

در پژوهش‌های پیشین نظیر [۳۴] و [۳۶]، پژوهشگران به‌طور عمده، تمامی ابعاد دادهٔ ورودی را مختل می‌کردند. در روش پپرنو و همکاران، بخشی از ابعاد ورودی برای مختل‌شدن انتخاب می‌شود. درواقع، ابعادی که دارای امتیاز بالاتری باشند، مختل می‌شوند. اگرچه محاسبهٔ امتیاز تأثیرگذاری برجستگی ابعاد داده زمان‌بر است، اما مزیت آن این است که تغییرات بسیار اندکی روی دادهٔ ورودی اعمال می‌شود.

در روش آن‌ها، نخست نقشهٔ برجستگی هر یک از عناصر[۱] ورودی توسط رابطهٔ (۱۳) محاسبه می‌شود.

$$S(x,t)[i] = \begin{cases} 0, & \frac{\partial F_t}{\partial x_i}(x) < 0 \ \ or \ \ \sum_{j \neq t} \frac{\partial F_j}{\partial x_i}(x) > 0 \\ \frac{\partial F_t}{\partial x_i}(x) \left| \sum_{j \neq t} \frac{\partial F_j}{\partial x_i}(x) \right|, & otherwise \end{cases} \quad (۱۳)$$

در رابطهٔ (۱۳)، منظور از $\boldsymbol{S(x,t)[i]}$ مقدار برجستگی عنصر **i**ام از داده **x** برای برچسب ردهٔ هدف **t** است. منظور از $\boldsymbol{F_t}$ خروجی تابع هزینه مبتنی‌بر ردهٔ هدف**t** است. همچنین، مقدار $\left[\frac{\partial F_j}{\partial x_i}\right]_{ij}$ توسط ماتریس ژاکوبین تابع هزینه **F** قابل محاسبه است.

سو[۲] و همکاران در [۲۷] از روش بهینه‌سازی تکامل تفاضلی[۳] برای به‌دست‌آوردن پاسخ مسئلهٔ پیداکردن اختلال بهینه استفاده کردند. آن‌ها در تلاش بودند تا در الگوریتم‌شان با تغییر یک پیکسل، برچسب ردهٔ ورودی را تغییر دهند. روند پیشروی آن‌ها نیز مبتنی‌بر اصول پردازش‌های تکاملی، تولید جمعیت و روال مشابه در فضای بهینه‌سازی تکاملی است. نظر سو و همکاران در دسته‌بندی فعالیت‌های اختلال محدود به نرم $\boldsymbol{l_0}$ قرار می‌گیرد. نکتهٔ بارز اهمیت روش آن‌ها این است که نیازی به تابع هزینه و یا اطلاعات الگو نیست و تنها نیاز به خروجی الگوی یادشده دارد (احتمال برچسب‌های پیش‌بینی شده). بنابراین این روش جزء حملات جعبه‌سیاه نیز قرار می‌گیرد.

روش *کارلینی و وگنر* C&W[4]: پپرنو و همکاران در پژوهش [۲۸]، خالص‌سازی دفاعی[5] را به‌عنوان راه حلی جهت مقاوم‌شدن در مقابل نمونه‌های خصمانه ارائه کردند. کارلینی و وگنر در [۱۶] به‌صورت ویژه به ارائهٔ سه رویکرد حمله در پاسخ به پژوهش پپرنو و همکاران [۲۸] پرداختند. رویکرد حمله پپرنو و همکاران، اختلال را با محدودیت نرم $\boldsymbol{l_0}$، $\boldsymbol{l_1}$ و $\boldsymbol{l_\infty}$ به‌دست‌می‌آورد. آن‌ها نشان دادند که روش خالص‌سازی دفاعی برای دفاع در مقابل نمونه‌های خصمانه مفید نیست. آن‌ها در نخست طرح کلی مسئلهٔ پیداکردن نمونهٔ خصمانه را در رابطهٔ (۱۲) بیان کرده، سپس، برای هر بخش از آن، معادل‌هایی را پیشنهاد دادند که درنهایت به پاسخ بهینه رابطه ختم شود.

در رابطهٔ (۱۲)، منظور از $\boldsymbol{Dist}$ معیار فاصله است[۶].

$$\begin{aligned} &min\ Dist(x, x+\eta) \\ &s.t.\ F(x+\eta) = y_{target} \\ &\quad x+\eta \in [0,1]^n \end{aligned} \quad (۱۲)$$

آن‌ها بیان کردند باتوجه‌به‌این‌که شرط $\boldsymbol{F(x+\eta) = y_{target}}$ غیرخطی است، باید حالت دیگری را برای در نظر گرفتن این شرط استفاده کرد. آن‌ها تلاش کردند تا تابعی معرفی نمایند که قابلیت جایگزینی با شرط یادشده را داشته‌باشد. بنابراین، آن‌ها تابعی که بیشترین شباهت خروجی به شرط مذکور را داشته‌باشد، $\boldsymbol{f}$ نامیدند و اقدام به جایگزین‌کردن آن با شرط مسئله کردند. به عبارت دیگر، شرط یادشده در صورتی صادق است که بتوان یک تابع مانند $\boldsymbol{f}$ پیدا کرد که در رابطهٔ (۱۴) صدق کند.

$$F(x+\eta) = y_{target} \iff f(x+\eta, t) \leq 0 \quad (۱۴)$$

سپس، تلاش کردند تا برای بخش‌های مختلف نیز به همین منوال معادل‌هایی را لحاظ نمایند و در نهایت به رابطهٔ (۱۵) رسیدند:

$$\begin{aligned} &\textbf{minimize} \left\| \frac{1}{2}(\tanh(w)+1) - x \right\|_2^2 + c \\ &\cdot loss_{f, y_{target}}\left(\frac{1}{2}(\tanh(w)+1)\right) \\ &s.t.\ \ f(\hat{x}) = \max(max\{[Z(\hat{x})]_i : i \neq t\} - [Z(\hat{x})]_t, -\kappa) \end{aligned} \quad (۱۵)$$

آن‌ها بیان کردند که برای حملات هدفمند، می‌توان از $\boldsymbol{\kappa = 0}$ استفاده کرد. همچنین، هرچه تمرکز، روی انتقال اختلال به شبکه‌های دیگر نیز مطرح باشد، می‌توان از مقادیر $\boldsymbol{\kappa}$ بالاتر استفاده کرد. آن‌ها بیان کردند که شبکهٔ

---

[۱] در داده‌های تصویری منظور از عنصر، پیکسل‌های تصویر است.

[2] Su

[3] Differential Evolution

[4] Carlini and Wanger

[5] Defensive distillation

[۶] همانند بیشتر روش‌هایی که تاکنون تحت عناوین نرم‌های مختلف $\boldsymbol{L_0}$، $\boldsymbol{L_2}$ و $\boldsymbol{L_\infty}$ به‌عنوان معیار فاصله بیان شد.

[DOI: 10.61186/jsdp.20.2.113]

خالص‌سازی‌شده در قبال حملهٔ l_2 مقاوم نبوده و قابل‌نفوذ است. منظور $\mathbf{Z}$ لاجیت (خروجی لایهٔ آخر شبکه قبل از اعمال **softmax**) و $\mathbf{w}$ متغیر جایگزین $\boldsymbol{\eta}$ در $\boldsymbol{Dist(x, x+\eta)}$ و رویکرد آن‌ها جزأ روش‌های تکرارشونده است.

روش اختلال خصمانهٔ فراگیر[1]: اغلب روش‌های حملهٔ بیان‌شده، اختلال را متناظر با هر نمونه داده به‌دست‌می‌آورند. موسوی دزفولی و همکاران در [۱۷]، تلاش کردند تا رویکردی جهت تولید اختلال خصمانه فراگیر ارائه کنند. با استفاده از اختلال خصمانهٔ فراگیر، به‌جای‌این‌که اختلال مدنظر به نمونه دادهٔ وابسته باشد، به شبکهٔ (الگو) وابسته است. درواقع، زمانی‌که اختلال فراگیر روی تعدادی از نمونه داده‌های ورودی اعمال شود، موجب می‌شود تا حملهٔ خصمانه با موفقیت انجام شود و ردهٔ نمونه‌های متناظر تغییر یابد. گفتنی است که این اختلال نیز باید شرط نامحسوس‌بودن را از دیدگاه انسانی داشته‌باشد، ولی آن‌ها نیز اشاره کردند که اختلال یادشده شبه‌نامحسوس[2] است. آن‌ها همچنین، اشاره کردند الزامی بر فراگیربودن اختلال برای همه تصاویر نبوده و تنها اختلالی مدنظر است که بیشینهٔ تصاویر را مختل کند و سامانه را فریب دهد. آن‌ها شروطی جهت پیداکردن اختلال کمینه معرفی کردند که به‌صورت روابط (۱۶) بیان شده‌است:

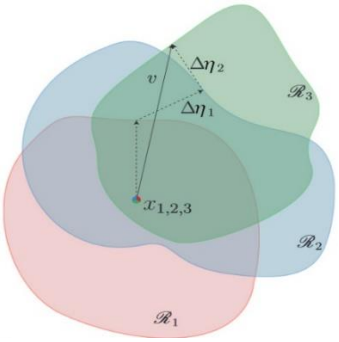
$$\|\eta\|_p \le \xi \qquad (۱۶)$$
$$\mathbb{P}_{x\sim X}\big(F(x+\eta) \neq F(x)\big) \ge 1-\delta$$

در روابط (۱۶)، منظور از $\mathbb{P}$ احتمال و منظور از کل عبارت، احتمال وجود اختلال روی تعدادی از نمونه‌ها نظیر $\boldsymbol{x}$ است. همچنین، منظور از $\boldsymbol{\xi}$ شاخص کنترل‌کنندهٔ میزان اختلال و $\boldsymbol{\delta}$ تعیین‌کنندهٔ نرخ فریب نمونه‌های نمونه‌برداری‌شده از توزیع $\boldsymbol{X}$ است. آن‌ها بیان کردند که اگر اختلال فراگیر روی مجموعه دادهٔ اندک محاسبه شود، نیاز است تا نرخ فریب، بزرگ انتخاب شود. گفتنی است که روش آن‌ها از روش‌های تکرارشونده بوده، و در هر تکرار کوچک‌ترین اختلال محاسبه شده و با مقدار جاری آن در آن مرحله، جمع بسته‌می‌شود. مسئلهٔ بهینه‌سازی آن‌ها به‌صورت رابطهٔ (۱۷) بیان شده‌است:

$$\Delta\eta \leftarrow arg\min_r \|r\|_2 \qquad (۱۷)$$
$$s.t. F(x_i+\eta+r) \neq F(x_i)$$

آن‌ها به‌منظور محدودکردن شرط رابطهٔ (۱۷)، اختلال به‌دست‌آمده را در هر تکرار روی کره $\boldsymbol{l_p}$ با مرکزیت صفر و شعاع $\boldsymbol{\xi}$ تصویر می‌کنند. به عبارت دیگر، این بیان در قالب رابطهٔ (۱۸) بیان می‌شود:

$$Err(X_\eta) = \frac{1}{m}\sum_{i=1}^{m} 1_{F(x_i+\eta)\neq F(x_i)} \ge 1-\delta \qquad (۱۸)$$

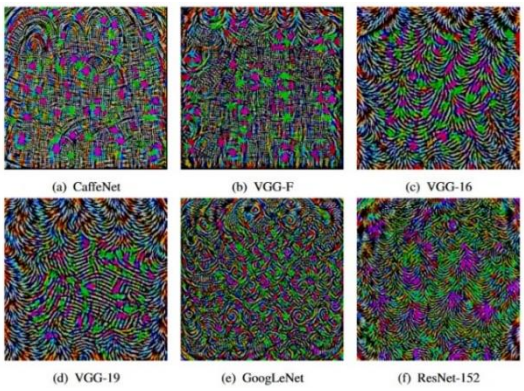


(شکل- ۶) طرح بازنمایی الگوریتم اختلال خصمانهٔ فراگیر [۱۷]

(Figure 7) Schematic representation of the universal adversarial perturbation algorithm [17]

در (شکل- ۶)، خلاصه‌ای از مراحل الگوریتم اختلال خصمانهٔ فراگیر به تصویر کشیده شده‌است. در این شکل نمونه‌داده‌های $\boldsymbol{x_1}$، $\boldsymbol{x_2}$ و $\boldsymbol{x_3}$ در نمای دوبعدی به‌طورتقریب، روی هم افتاده‌اند. محدوده‌های هر دسته‌بند با استفاده از هاله‌های رنگی $\boldsymbol{\mathfrak{R}_1}$، $\boldsymbol{\mathfrak{R}_2}$ و $\boldsymbol{\mathfrak{R}_3}$ نشان داده شده‌است.

الگوریتم روش آن‌ها، در هر تکرار، اختلال کمینه را تجمیع کرده و سعی می‌کند دادهٔ مختل‌شده را به خارج از محدوده $\boldsymbol{\mathfrak{R}_i}$ هدایت کند. همین روال به‌صورت تکرارشونده تا خارج‌شدن از نواحی مختلف و به میزانی که کاربر درخواست کرده، ادامه می‌یابد.



(شکل- ۷)- نمونه‌ای از اختلال‌های فراگیر محاسبه‌شده برای معماری‌های مختلف شبکه‌های عصبی [۱۷] (برای نمایش بهتر مقادیر پیکسل‌ها تغییر مقیاس شده‌اند.)

(Figure 8) An example of universal perturbations computed for different neural network architectures (The pixel values are scaled for better performing the visibility) [17]

در (شکل- ۷) نمونه‌ای از اختلال‌های فراگیر محاسبه‌شده برای معماری‌های مختلف شبکه‌های عصبی ترسیم شده‌است. در بسیاری از این اختلال‌ها، الگوهای قابل‌توجهی مشاهده می‌شود. از طرفی در شبکه‌های مختلف الگوهای مشابه نیز مشاهده می‌شود.

[1] Universal Adversarial Perturbation
[2] Quasi imperceptible

[DOI: 10.61186/jsdp.20.2.113]

حملهٔ جعبه‌سیاه با استفاده از بهینه‌سازی مرتبه صفر: چن[1] و همکاران در [۲۹] رویکردی مستقل از گرادیان ارائه کردند. آن‌ها برخلاف بیشتر روش‌های تولید نمونهٔ خصمانه که اغلب با گرادیان کار می‌کنند، روشی ارائه کردند که نیازی به استفاده ازگرادیان نداشته‌باشد. در نتیجه، روش آن‌ها گزینهٔ مناسبی برای استفاده در روش‌های جعبه سیاه به شمار می‌رود. آن‌ها نظر خود را از پژوهش کارلینی و وگنر [۱۶] اقتباس کردند. درواقع، آن‌ها با ترکیب تابع $\boldsymbol{f}$ مطرح‌شده در پژوهش کارلینی و وگنر [۳۰] به رابطهٔ (۱۹) دست یافتند.

$$\boldsymbol{f(\hat{x}) = \max(max\{log[Z(\hat{x})]_i : i \neq t\} - log[Z(\hat{x})]_{y_{target}}, -\kappa)} \quad (۱۹)$$

گفتنی است که آن‌ها با استفاده از مشتق متقارن، تخمینی از گرادیان نیز به‌صورت رابطهٔ (۲۰) به‌دست‌آورده‌اند:

$$\frac{\partial f(x)}{\partial x_i} \approx \frac{f(x + he_i) - f(x - he_i)}{2h} \quad (۲۰)$$

منظور از $\boldsymbol{h}$ در رابطهٔ (۲۰)، عددی ثابت با مقدار پیشنهادی ۰.۰۰۰۱ است. $\boldsymbol{e_i}$ نیز بردار پایهٔ استاندارد است که فقط درایه iامین عنصر آن یک است. تخمین خطای آن‌ها در مرتبه $\boldsymbol{O(h^2)}$ است. ازآنجاکه در روش **FGSM** از تابع علامت گرادیان استفاده می‌شود، چن و همکاران بر این عقیده بودند که نیازی به در اختیار داشتن مقدار دقیق گرادیان نبوده‌است؛ بنابراین، نظر مطرح‌شدهٔ آن‌ها نیز بر همین اصل بنا نهاده‌شده که با استفاده از تقریب و تابع واسط روال کار پیش برده‌شود. گفتنی است آن‌ها به‌منظور ارزیابی بهتر، تخمین گرادیان را چندین بار تکرار می‌کنند. تخمین هسین[2] مطرح‌شده در پژوهش‌آن‌ها نیز به‌صورت رابطهٔ (۲۲) است.

$$\frac{\partial^2 f(x)}{\partial x_{ii}^2} \approx \frac{f(x + he_i) - 2f(x) + f(x - he_i)}{h^2} \quad (۲۲)$$

آن‌ها از نام **ZOO** [3] برای پژوهش خویش استفاده کردند. در روش **ZOO**، نیازی به شاخص‌های الگو نبوده و با استفاده از تخمین گرادیان مراحل آتی آن پیش می‌رود. البته محاسبهٔ تخمین گرادیان امری زمان‌بر است. نتایج نشان داده‌است که روش آن‌ها از روش کارلینی و وگنر [۱۶] نیز بهتر عمل می‌کند.

[1] Chen
[2] Hessian
[3] Zero Order Optimization

تولید نمونهٔ خصمانه طبیعی[4]: *ژائو*[5] و همکاران در [۳۱] تلاش کردند تا با استفاده از چارچوب شبکهٔ تقابلی مولد[6] نمونهٔ خصمانه تصویری و متنی تولید کنند. نمونه‌های تولیدشده توسط پژوهش آن‌ها برای انسان نیز طبیعی به نظر می‌رسد. آن‌ها با استفاده از آموزش شبکهٔ تقابلی مولّد با تابع هزینه واسرشتین[7] به نام شبکهٔ تقابلی مولد واسرشتین (**WGAN**[8]) [۵۴] و [۵۵] این فعالیت را انجام دادند. همچنین، آن‌ها یک شبکه معکوس[9] $\boldsymbol{\mathcal{I}}$ را در کنار شبکهٔ تقابلی مولّد اصلی قرار دادند تا نمونه‌های خصمانه را به فضای نهان هدایت کند. طبق بیان آن‌ها خطای بازسازی نمونه $\boldsymbol{x}$ همراه با اختلاف بین توزیع گوسی $\boldsymbol{z}$ و $\boldsymbol{\mathcal{I}_\gamma(\mathcal{G}_\theta(z))}$، توسط تابع هزینه شبکهٔ معکوس ترکیب می‌شود. رسیدن به نقطهٔ بهینه و پایدار در شبکه‌های تقابلی مولد، دشوار است. برای بهینه‌شدن وضعیت پاسخ، *آرژوفسکی*[10] و همکاران در [۳۲] روش بهتری ارائه کردند که یادگیری را راحت‌تر و از رخداد فروپاشی حالت[11] در شبکهٔ تقابلی مولّد جلوگیری می‌کند. ژائو و همکاران باتوجه‌به بهبودهای متنوعی که پژوهش‌های مختلف، نظیر [۳۲] و [۳۴]–[۳۶] در یادگیری شبکهٔ تقابلی مولد ایجاد کردند، تابع هزینه‌ای به‌صورت رابطهٔ (۲۱) ارائه کردند:

$$\min_{\theta} \max_{\omega} \mathbb{E}_{x\sim p_x}[\mathcal{C}_\omega(x)] - \boldsymbol{\lambda} \cdot \mathbb{E}_{z\sim p_z}[\mathcal{C}_\omega(\mathcal{G}_\theta(z))] \quad (۲۱)$$

در رابطهٔ (۲۱) منظور از $\boldsymbol{\mathcal{C}_\omega}$ تابع منتقد[12] است. همچنین، از مقدارهای ۰.۱ و ۱ به‌ترتیب، برای حوزهٔ تصویر و متن جهت تخصیص به شاخص $\boldsymbol{\lambda}$ در رابطهٔ (۲۱) استفاده کردند.

ژائو و همکاران به‌منظور تولید نمونهٔ خصمانه، نخست یک شبکهٔ WGAN را روی مجموعه دادهٔ اصلی $\boldsymbol{X}$، آموزش می‌دهند. شبکهٔ یادشده شامل شبکه مولد $\boldsymbol{\mathcal{G}_\theta}$ بوده و وظیفهٔ نگاشت بازنمایی بردار $\boldsymbol{z \in \mathbb{R}^d}$ به نمونه‌های تولیدشده نظیر $\boldsymbol{x}$ از دامنه $\boldsymbol{X}$ را داراست. آن‌ها شبکهٔ معکوس $\boldsymbol{\mathcal{I}}$ که وظیفهٔ نگاشت نمونه‌های داده به بازنمایی $\boldsymbol{z}$ را داراست، جداگانه آموزش می‌دهند. در نهایت، در روند شبکهٔ تقابلی مولد، اختلاف بازنمایی تولیدشده با خروجی

[4] Generating Natural Adversarial Example
[5] Zhao
[6] Generative Adversarial Network
[7] Wasserstein
[8] Wasserstein Generative Adversarial Network
[9] منظور از معکوس بودن شبکه این است که از فضای داده به فضای نهان قابلیتِ نگاشتِ داده داشته‌باشد.
[10] Arjovsky
[11] Mode collapse
[12] Critic

[DOI: 10.61186/jsdp.20.2.113]

شبکه معکوس مقایسه می‌شود. درواقع، باید اختلاف بین بازنمایی تولیدشده و خروجی شبکه معکوس کمینه شود. مسئله بهینه‌سازی آن‌ها به‌صورت رابطهٔ (۲۳) است:

$$\hat{\boldsymbol{x}} = \boldsymbol{\mathcal{G}_\theta}(\hat{\boldsymbol{z}})$$
$$\hat{\boldsymbol{z}} = \underset{\tilde{\boldsymbol{z}}}{\mathbf{argmin}} \|\tilde{\boldsymbol{z}} - \boldsymbol{\mathcal{I}_\gamma}(\boldsymbol{x})\| \qquad (۲۳)$$
$$\boldsymbol{s.t.}\ \boldsymbol{F}(\boldsymbol{\mathcal{G}_\theta}(\tilde{\boldsymbol{z}})) \neq \boldsymbol{F}(\boldsymbol{x})$$

حمله تک‌پیکسل[1]: سو[2] و همکاران در [۲۷] تلاش کردند تا با تغییر دادن یک پیکسل، نمونهٔ خصمانه تولید نمایند. آن‌ها تصویر ورودی را به‌صورت برداری در نظر گرفتند که هر عنصر آن نشان‌دهندهٔ یک پیکسل از تصویر است. مسئلهٔ بهینه‌سازی پیداکردن اختلال در نگاه آنان به‌صورت رابطهٔ (۲۴) بیان می‌شود:

$$\min_{\hat{\boldsymbol{\eta}}} \boldsymbol{F}(\boldsymbol{x} + \boldsymbol{\eta}) \qquad (۲۴)$$
$$\boldsymbol{s.t}\ \|\boldsymbol{\eta}\|_0 \leq \boldsymbol{d}$$

در رابطهٔ (۲۴) منظور از ثابت **d**، عددی کوچک و این عدد برای حمله تک‌پیکسل عدد یک است. درواقع، منظور این است که تنها در $\boldsymbol{d}$ بعد می‌توان تغییرات ایجاد کرد و در دیگر ابعاد تغییری ایجاد نمی‌شود. آن‌ها به‌منظور پیداکردن پاسخ بهینه از تکامل تفاضلی[3] استفاده کردند. برتری استفاده از تکامل تفاضلی عدم نیاز به گرادیان شبکهٔ عصبی است. همچنین، از این رویکرد می‌توان برای توابع هزینه غیرمشتق‌پذیر استفاده کرد. در

۳- دفاع در برابر نمونه خصمانه

در بخش‌های قبل بیان شد که شبکه‌های عصبی عمیق نسبت‌به نمونه‌های خصمانه آسیب‌پذیر هستند و نیاز است بر مقاوم‌سازی آن‌ها تمرکز شود. در ادامه، انواع روش‌های مقابله با نمونه‌های خصمانه و رویکردهای دفاع در مقابل آن‌ها را مرور خواهیم‌کرد.

تعریف دفاع

به فرآیندهای مقابله با آسیب‌پذیری شبکه‌های عصبی عمیق، نسبت‌به نمونه‌های خصمانه دفاع گفته‌می‌شود. این فرآیندها شامل روش‌های متنوعی هستند که پاسخ الگو را نسبت‌به ورودی خصمانه کنترل می‌کنند. همچنین، در برخی روش‌ها دقت و تعمیم‌پذیری الگو نیز افزایش می‌یابد.

(جدول-۱)) مقایسهٔ روش‌های مختلف تولید نمونهٔ خصمانه انجام شده‌است.

[1] One-pixel attack
[2] Su
[3] Differential evolution

۳- دفاع[4] در برابر نمونه خصمانه

در بخش‌های قبل بیان شد که شبکه‌های عصبی عمیق نسبت‌به نمونه‌های خصمانه آسیب‌پذیر هستند و نیاز است بر مقاوم‌سازی آن‌ها تمرکز شود. در ادامه، انواع روش‌های مقابله با نمونه‌های خصمانه و رویکردهای دفاع در مقابل آن‌ها را مرور خواهیم‌کرد.

تعریف دفاع

به فرآیندهای مقابله با آسیب‌پذیری شبکه‌های عصبی عمیق، نسبت‌به نمونه‌های خصمانه دفاع گفته‌می‌شود. این فرآیندها شامل روش‌های متنوعی هستند که پاسخ الگو را نسبت‌به ورودی خصمانه کنترل می‌کنند. همچنین، در برخی روش‌ها دقت و تعمیم‌پذیری الگو نیز افزایش می‌یابد.

(جدول-۱) طبقه‌بندی روش‌های حملهٔ پرداخته‌شده در این مقاله. T بیانگر هدفمندبودن رویکرد مدنظر، W و B به ترتیب روش‌های جعبه‌سفید و جعبه‌سیاه است.

Different categories of the selected attack methods in this paper. T indicates the purposefulness of the approach, and W and B are White-box and Black-box approaches, respectively.

روش حمله	نوع	تعداد مرحله	نرم اختلال	همگانی	وضعیت هدفمند
L-BFGS [۹]	W	تک	∞	–	T
FGSM [۸]	W	تک	∞	–	T
IFGSM/BIM [۱۱]	W	چند	∞	–	-
ILCM [۱۱]	W	چند	∞	–	T
Rep. Adversary [۲۵]	W	تک	∞	–	T
JBM [۲۶]	W	تک	0	–	T
C & W [۱۶]	W	چند	0, 2, ∞	–	T
DeepFool [۱۵]	W	چند	2, ∞	–	-
UAP [۱۷]	W	چند	2, ∞	بله	-
ZOO [۲۹]	B	چند	ترکیبی	–	-
Natural GAN [۳۱]	B	تک	0	–	-
One-Pixel [۲۷]	B	چند	0	–	-

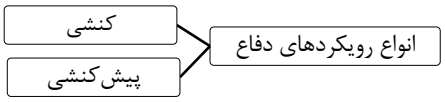
انواع روش‌های دفاع

از نگاه کلان مطابق (شکل۸)، روش‌های دفاع و مقابله با نمونه‌های خصمانه به دو دسته روش‌های کنشی[5] و روش‌های پیش‌کنشی[6] تقسیم‌بندی می‌شوند [۳۷]. در روش‌های کنشی پس از ساخت شبکه عصبی (الگو)، درصدد بررسی آسیب‌پذیری و مقابله با نمونهٔ خصمانه برمی‌آییم. در برخی روش‌ها پس از تشخیص خصمانه‌بودن نمونهٔ ورودی، آن ورودی از ادامه فرآیند

[4] Defense
[5] Reactive
[6] Preactive

[DOI: 10.61186/jsdp.20.2.113]

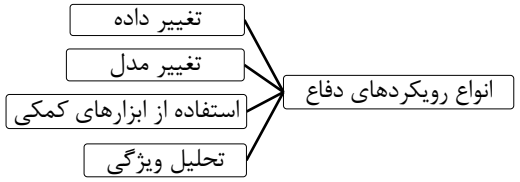
خارج شده، پس از طی چند مرحله (مراحل پاک‌سازی) دوباره به فرآیند عادی برگشت داده‌می‌شود. در روش‌های پیش‌کنشی بیشتر به ماهیت شبکهٔ عصبی عمیق توجه می‌شود. درواقع، سعی می‌شود تا طراحی معماری شبکه به‌گونه‌ای انجام پذیرد که این معماری نسبت‌به نمونه‌های متنوع خصمانه مقاوم باشد.



(شکل۸) - انواع رویکردهای دفاع بر اساس نحوهٔ واکنش در مقابل نمونه‌های خصمانه

(Figure 9) Different kinds of defense approaches based on their reaction against adversarial examples

از دیدگاهی دیگر می‌توان روش‌های مقابله با نمونه‌های خصمانه را به چهار دستهٔ تغییرِ داده[1]، تغییر الگو[2] و استفاده از ابزارهای کمکی[3] و تحلیل فضای ویژگی تقسیم‌بندی کرد. همچنین، در برخی پژوهش‌های اخیر نظیر [۲۴] و [۴۲]–[۴۴] به میزان درهم‌تنیدگی فضای ویژگی یادگرفته‌شده نیز اشاره شده‌است. در (شکل- ۹) تقسیم‌بندی یادشده مشاهده می‌شود.



(شکل- ۹) - انواع رویکردهای دفاع بر اساس اقدامات عملیاتی و تحلیلی در مقابل نمونه‌های خصمانه

(Figure 10) Different kinds of defense approaches based on operational and analysis actions against adversarial examples

در ادامه مروری بر برخی از پژوهش‌ها خواهیم‌داشت:

تغییر داده

در این روش به‌منظور مقابله با نمونه‌های خصمانه، تغییرات و تنوعات هدفمند فراوانی در مجموعه داده‌های آموزشی ایجاد و این کار موجب افزایش حجم مجموعه دادگان می‌شود. این امر به این علت انجام می‌پذیرد که الگو پس از مرحلهٔ آموزش تنوع بیشتری از داده را مشاهده کرده و درنتیجه، مقاوم‌تر می‌شود. در برخی از روش‌ها، فرآیندی به مرحله آزمون اضافه شده و موجب حذف، یا ترمیم داده می‌شود. این نوع از تغییر داده نیز در این دسته قرار می‌گیرد. در ادامه سعی شده به برخی از روش‌های مطرح در این دسته‌بندی اشاره شود.

آموزش خصمانه[4]: در این روش نمونه‌های خصمانه به مجموعه دادگان آموزشی اضافه می‌شوند. به بیان دیگر نمونه‌های خصمانه از این پس به‌عنوان نمونه‌های مجاز[5] محسوب و در کنار مجموعه دادگان آموزشی در آموزش استفاده می‌شوند. سپس، الگو با مجموعهٔ دادگان جدید آموزش می‌بیند و همین امر موجب می‌شود تا نسبت‌به حالت پیشین، نسبت‌به حملات خصمانه مقاوم‌تر عمل کند. البته این موضوع جای بحث دارد و نمی‌توان بیان کرد که نسبت‌به همه نمونه‌های خصمانه مقاوم خواهدبود. زیرا نمونه‌های خصمانه با روش‌های متنوعی تولید می‌شوند. همان‌طورکه بیان شد، در پژوهش‌های گوناگونی از آموزش خصمانه به اضافه‌کردن نمونه‌های خصمانه[6] به مجموعه دادگان اصلی و افزایش مجموعه دادگان و آموزش دوبارهٔ الگو روی آن‌ها یاد کرده‌اند.

سژدی و همکاران در [۹] به اضافه‌کردن نمونه‌های خصمانه و در نتیجه، افزایش مقاوم‌سازی الگو نسبت‌به نمونه‌های خصمانه اشاره کردند. این بحث توسط گودفلو و همکاران در مقاله [۸] بسط داده شده و جوانب مختلفی از آن پوشش داده‌شد. آن‌ها از آن به نام تابع هزینه خصمانه[7] یاد و بیان کردند که می‌توان با اضافه‌کردن یک عبارت منظم‌ساز[8] به تابع هزینه، تأثیری به‌نسبت برابر با افزودن نمونه‌های خصمانه به مجموعه داده ایجاد کرد. امروزه اصطلاح «آموزش خصمانه» اغلب به اضافه‌کردن این عبارت منظم‌ساز مبتنی‌بر حمله به تابع هزینهٔ اصلی اطلاق می‌شود. برای مثال، آن‌ها تابع هزینهٔ خصمانه را برای روش حمله **FGSM** به‌صورت ترکیب خطی از تابع هزینه و عبارت منظم‌ساز همانند رابطه (۲۵) بیان کرده‌اند. به عبارت منظم‌ساز اضافه‌شده، عبارت خصمانه[9] نیز گفته می‌شود.

$$J^*(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\alpha} J(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y}) + (\mathbf{1} - \boldsymbol{\alpha})\, J(\boldsymbol{\theta}, \boldsymbol{x} + \boldsymbol{\varepsilon sign}\big(\nabla_x J(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y})\big), \boldsymbol{y}) \quad (۲۵)$$

در رابطهٔ (۲۵) منظور از $\boldsymbol{J}$ تابع هزینه، $\boldsymbol{J}^*$ تابع هزینه بهینه‌شده، $\boldsymbol{\theta}$ شاخصه‌های الگو (وزن‌های شبکه عصبی)، $\boldsymbol{x}$ دادهٔ آموزشی، $\boldsymbol{y}$ برچسب دادهٔ آموزشی و $\boldsymbol{\alpha}$ شاخص ثابت منظم‌ساز با مقدار پیشنهادی $\boldsymbol{\alpha} = \mathbf{0.5}$ است [۸]. تعریف تابع هزینه به‌صورت رابطهٔ (۲۵)، موجب می‌شود تا عبارت

[1] Modifying data
[2] Modifying model
[3] Auxiliary tools

[4] Adversarial training
[5] Legalized
[6] نمونه‌های خصمانه هر مرحله از یادگیری مدل
[7] Adversarial cost function
[8] Regularization term
[9] Adversarial term

[DOI: 10.61186/jsdp.20.2.113]

منظم‌ساز به‌عنوان عبارت جریمه‌کننده لحاظ شود و با تابع هزینهٔ اصلی جمع شود. در نهایت، وضعیت بهینه زمانی خواهدبود که هر دو عبارت دخیل در تابع هزینهٔ کلی کمینه شوند (تابع هزینه قبلی و عبارت منظم‌ساز جدید).

عبارت منظم‌ساز آموزش خصمانهٔ رویکرد سژدی و همکاران، مبتنی‌بر حمله **FGSM** است. استفاده از حمله **PGD** در آموزش خصمانه توسط مادری و همکاران در [۲۴] پیشنهاد شده و دارای هزینهٔ محاسباتی بالایی نسبت‌به **FGSM** است. *ونگ*[1] و همکاران در [۴۱] با افزودن یک مرحلهٔ اولیه (مرحله مقداردهی تصادفی اختلال)، عملکرد استفاده از **FGSM** در آموزش خصمانه را تا اندازهٔ استفاده از **PGD** در آموزش خصمانه افزایش دادند و آن را **FFGSM**[2] نامیدند. این امر موجب افزایش سرعت آموزش و کاهش حجم محاسبات شده‌است. گفتنی است روش **FFGSM** به‌صورت مستقل دارای یک روش حمله و همانند سایر حملات، با استفاده از روش آموزش خصمانه قابل استفاده برای دفاع از حملات است؛ بنابراین، در بخش آزمایش‌ها از این نوع حمله برای بررسی بهبودهای آن نسبت‌به **FGSM** نیز استفاده شده‌است.

آموزش خصمانهٔ جمعی[3]: *ترامر*[4] و همکاران در [۴۲]، نظر آموزش خصمانهٔ جمعی را مطرح کردند. آموزش خصمانهٔ جمعی، طیف بیشتری از نمونه‌های خصمانه را پوشش می‌دهد و روال عملکرد آن توسعهٔ آموزش خصمانه است. این امر موجب شده تا نسبت‌به آموزش خصمانهٔ عادی بهتر عمل کند و نسبت‌به نمونه‌های خصمانهٔ بیشتری مقاوم باشد. آن‌ها بیان کردند که آموزش خصمانه به کمینهٔ سراسری نزدیک می‌شود و در نتیجه، نسبت‌به حملات جعبه‌سیاه الزاماً مقاوم نیست. درواقع، آن‌ها اصل انتقال‌پذیری را که گاهی در حملات جعبه‌سیاه مطرح می‌شود، تأیید و بیان کردند که این اصل در آموزش خصمانه نیز برقرار است. آن‌ها با استفاده از تلفیق یادگیری جمعی و آموزش خصمانه، رویکرد آموزش خصمانهٔ جمعی را ارائه کردند. همچنین، بیان کردند که آموزش خصمانهٔ جمعی به تولید اختلال‌های متنوع منجر شده و در نتیجه، افزایش تنوع داده‌های آموزشی را به همراه دارد؛ بنابراین، منطقی است که نسبت‌به نمونه‌های خصمانهٔ بیشتری مقاوم باشد. همچنین، آن‌ها بیان کردند که آموزش خصمانه در داده با ابعاد بزرگ کارکرد مناسبی ندارد و به عبارت دیگر، تأثیر کمتری در مقاوم‌بودن دارد. آموزش خصمانهٔ جمعی، به‌گونه‌ای نسخهٔ توسعه‌یافته آموزش خصمانه است؛ بنابراین، از طریق الگوهای از پیش‌تعلیم‌دیده‌شدهٔ موجود، نمونهٔ خصمانه تولید و در نتیجه، موجب تکمیل مجموعه داده خصمانهٔ جمعی در آموزش کلی می‌شود. الگوی اصلی نیز می‌تواند از این مجموعه داده جدید استفاده کند و تنوع مختلفی از نمونه‌های خصمانه را (که توسط الگوهای مختلف در آن گروه تولید شده) استفاده کند و الگو نسبت‌به نمونه‌های خصمانهٔ بیشتری مقاوم باشد.

پنهان‌سازی گرادیان[5]: ترامر و همکاران در پژوهش [۴۲] نیز رویکردی جهت مقابله با روش‌های حملهٔ مبتنی‌بر گرادیان ارائه کردند. این رویکرد دارای قابلیت دفاع در مقابل روش‌های حمله مبتنی‌بر گرادیان نظیر **FGSM** و نسخه‌های توسعه‌یافتهٔ آن است. آن‌ها بیان کردند که می‌توان الگوهایی ارائه کرد که از گرادیان استفاده نشود. برای مثال الگوهای غیرمشتق‌پذیر نظیر درخت تصمیم، نزدیک‌ترین همسایه یا درخت تصادفی فاقد گرادیان هستند و در نتیجه، با روش‌های حمله مبتنی‌بر گرادیان به‌راحتی مقابله می‌کنند.

پپرنو و همکاران در [۴۳] رویکردی ارائه کردند که موجب نقض روش دفاع «پنهان‌سازی گرادیان» شد. درواقع، آن‌ها پیشنهاد تعلیم یک دسته‌بند جایگزین[6] با نتایج مشابه را دادند. سپس، با استفاده از گرادیان الگوی جایگزین، تولید نمونهٔ خصمانه صورت می‌گیرد. در نتیجه، طبق اصل انتقال‌پذیری نمونه خصمانه، از آن نمونه‌ها برای حمله به الگوهای بدون گرادیان استفاده کردند. درواقع، آن‌ها موفق به شکست روش پنهان‌سازی گرادیان (به‌عنوان یک روش دفاع در مقابل نمونه‌های خصمانه) شدند.

منع انتقال‌پذیری[7]: خاصیت انتقال‌پذیری جزء ویژگی‌هایی است که با تغییر ماهیت شبکه‌ها همچنان در برخی موارد مشاهده می‌شود. باتوجه‌به‌این‌که از انتقال‌پذیری نمونه‌های خصمانه در حملات جعبه‌سیاه به‌وفور استفاده می‌شود، حسینی و همکاران در [۴۴] رویکردی جهت کاهش انتقال‌پذیری معرفی کردند. آن‌ها با معرفی روش سه مرحله‌ای برچسب‌گذاری پوچ[8]، موفق به کاهش انتقال‌پذیری از یک شبکه به شبکهٔ دیگر شدند. نظر آنها اضافه‌شدن برچسب پوچ به مجموعه برچسب‌های خروجی شبکه است. روش آن‌ها با دیدن نمونهٔ خصمانه به‌جای افزایش احتمال برچسب‌های اصلی، احتمال

[1] Wang
[2] Fast adversarial training using FGSM (FFGSM)
[3] Ensemble adversarial training
[4] Tramèr
[5] Gradient hiding
[6] Surrogate
[7] Blocking the transferability
[8] Null labeling

[DOI: 10.61186/jsdp.20.2.113]

برچسب پوچ را افزایش داد. این امر موجب می‌شود تا روش آن‌ها نسبت‌به خیلی از روش‌های مرسوم حمله مقاوم باشد. آن‌ها درنهایت، بیان کردند که برچسب پوچ تأثیری در میزان صحت عملکرد دسته‌بند ایجاد نمی‌کند.

فشرده‌سازی داده[1]: *زیوگیت*[2] و همکاران در [۴۵] تلاش کردند تا تأثیر فشرده‌سازی فرمت **JPG** را روی نمونه‌های خصمانه بررسی کنند. آن‌ها دریافتند که فشرده‌سازی فرمت **JPG**، روی بهبود صحت الگو در حملات مبتنی‌بر گرادیان نظیر **FGSM** تأثیر به‌سزایی دارد. *دس*[3] و همکاران نیز در [۴۶] از رویکردی مشابه جهت فشرده‌سازی فرمت **JPEG** استفاده کردند. دس و همکاران آن را به‌عنوان راه‌حلی جهت دفاع در مقابل حملات مبتنی‌بر گرادیان نظیر **FGSM** و **DeepFool** بیان کردند. گفتنی است که رویکرد دس و همکاران و همچنین، زیوگیت و همکاران در مقابل حملات قدرتمند نظیر روش کارلینی وگنر [۱۶] چندان مؤثر نیستند.

اختر و همکاران در [۴۷] به ارائهٔ الگویی جهت دفاع نسبت‌به حملات خصمانه فراگیر پرداختند. در روش آن‌ها یک مرحله قبل از ارسال داده به الگو اضافه شده‌است. در مرحلهٔ تعبیه‌شدهٔ آن‌ها، فرایند اصلاح[4] داده انجام می‌شود. در روش آن‌ها از تشخیص‌دهندهٔ اختلال استفاده شده که از تبدیل کسینوسی گسسته[5] که اغلب در فشرده‌سازی نیز استفاده می‌شود، بهره گرفته شده‌است. مشکل عمدهٔ رویکردهای دفاع مبتنی‌بر روش‌های فشرده‌سازی این است که موجب کاهش صحت روی داده‌های اصلی و حالت اولیه می‌شوند. همچنین، رویکردهایی که فشرده‌سازی اندکی دارند نیز، در اغلب موارد در قبال روش‌های حمله قدرتمند، مقاوم نیستند.

تصادفی‌سازی داده[6]: *شی*[7] و همکاران در [۴۸] تلاش کردند تا تأثیر تغییر اندازهٔ دادهٔ خصمانه در دسته‌بند را بررسی کنند. آن‌ها نشان دادند که تغییر اندازهٔ تصادفی موجب کاهش تأثیرگذاری نمونهٔ خصمانه بر الگو می‌شود. همچنین، بیان کردند که قراردادن بافت[8]های تصادفی در تصاویر موجب کاهش تأثیرگذاری نمونه‌های خصمانه می‌شود. آن‌ها با تعریف یک ماژول[9] در راستای افزودن یک مرحله پیش‌پردازش غیرمشتق‌پذیر به نام تعبیه محلی‌خطی[10] توانستند بر اختلال موجود در نمونه‌های خصمانه غالب شوند. گفتنی است که برای جامعیت بهتر، آن‌ها در مرحلهٔ آموزش نیز از آن ماژول استفاده کرده‌اند. همچنین، پس از مرحلهٔ عبور از ماژول، نوفهٔ تصادفی گوسی نیز به آن افزوده می‌شود که این امر بهبوددهندهٔ مقاومت الگوست.

تفکیک ویژگی‌های مقاوم از غیرمقاوم: *الیاس*[11] و همکاران در [۴۹]، نگرش دیگری به مسئلهٔ وجود نمونه‌های خصمانه ارائه کردند. آن‌ها بر این نظر بودند که وجود نمونه‌های خصمانه از ماهیت داده است و ارتباط کمتری به الگو دارد. بنابراین، آن‌ها توصیه کرده‌اند که احتمالاً باید مجموعهٔ دادگان را غنی‌تر کرد تا شبکه بر چنین مشکلاتی غالب شود. آن‌ها بر این باور بودند که شبکه‌های عصبی به ویژگی‌هایی که از دید عامل انسانی اهمیت ندارند، حساس می‌شوند. برای مثال، از دید عامل انسانی، این‌که موهای بدن یک گربه الگوی افقی دارد و موهای بدن یک سگ الگوی خطوط عمودی دارد، چندان اهمیت بالایی ندارد. اما از نگاه عامل انسانی شکل ظاهری اجزای این دو حیوان نظیر گوش، صورت و... جزء علائم مؤثّر در تشخیص آنها هستند (که شبکه‌های عصبی بعضاً از آن‌ها غافلند). آن‌ها با تفکیک ویژگی‌های داده به سه دسته ویژگی‌های $\boldsymbol{\rho}$-مفید، ویژگی‌های $\boldsymbol{\gamma}$-مقاوم و ویژگی‌های مفید غیرمقاوم، سعی کردند مسئله را با بیانی ریاضی بازنویسی نمایند. زمانی یک ویژگی $\boldsymbol{\rho}$-مفید خواهدبود که همبستگی بین آن ویژگی‌ها، با برچسب نهایی در بالاترین سطح ممکن قرارگیرد. درواقع، اگر مجموعه دادگان از توزیع داده شده $\boldsymbol{D}$ نمونه‌برداری شوند، ویژگی‌های $\boldsymbol{\rho}$-مفید به‌صورت رابطهٔ (۲۶) قابل بیان هستند:

$$\mathbb{E}_{(x,y)\sim D}[y \cdot f(x)] \geq \rho \qquad (۲۶)$$

همان‌طورکه در رابطهٔ (۲۶) آورده شده، از عملیات ضرب برای بیان همبستگی ویژگی‌ها و برچسب خروجی استفاده شده‌است (در این مثال برچسب $\pm\mathbf{1}$ است). هرچه این میزان عدد بزرگ‌تر و مثبتی باشد، نشان‌دهندهٔ این است که ویژگی مفیدتری بوده است. ویژگی‌های $\boldsymbol{\gamma}$-مقاوم در زمان وجود نمونهٔ خصمانه همچنان با برچسب خروجی همبسته هستند. بیان ریاضی این ویژگی‌ها به‌صورت رابطهٔ (۲۷) بیان شده‌است:

$$\mathbb{E}_{(x,y)\sim D}\left[\inf_{\delta\in\eta} y \cdot f(x+\delta)\right] \geq \gamma \qquad (۲۷)$$

[1] Data compression
[2] Dziugaite
[3] Das
[4] Rectifying
[5] Discrete Cosine Transform
[6] Data randomization
[7] Xie
[8] Texture
[9] Module
[10] Linear Locally Embedding (LLE)
[11] Ilyas

[DOI: 10.61186/jsdp.20.2.113]

همان‌طورکه در رابطهٔ (۲۷) نیز آورده‌شده، کمترین اختلالی که زیر مجموعهٔ اختلال‌های خصمانه باشد و موجب افزایش همبستگی ویژگی‌های ورودی و برچسب شود، و از حد آستانه $\boldsymbol{\gamma}$ بیشتر باشد، از ویژگی‌های $\boldsymbol{\gamma}$-مقاوم به شمار می‌رود. ویژگی‌های مفید و غیرمقاوم، ویژگی‌های $\boldsymbol{\rho}$-مفیدی هستند که به‌ازای هیچ مقدار $\boldsymbol{\gamma} \geq \mathbf{0}$، $\boldsymbol{\gamma}$-مقاوم نیستند. این ویژگی‌ها موجب افزایش صحّت دسته‌بندی در شرایط غیرخصمانه هستند، اما زمانی‌که شرایط خصمانه شود، این ویژگی‌ها به‌علت تأثیر در کاهش همبستگی مناسب نیستند. الیاس و همکاران طبق تعاریف بیان‌شده تلاش کردند تا ویژگی‌های مقاوم از سایر ویژگی‌ها جدا شوند و تنها روی آن ویژگی‌ها آموزش انجام شود. آن‌ها بر این باور بودند که رویکرد آن‌ها علاوه بر افزایش صحت، موجب افزایش مقاومت‌پذیری نیز می‌شود. برای این امر آن‌ها تلاش کردند تا ویژگی‌های مقاوم را به‌صورت رابطهٔ (۲۸) استخراج نمایند.

$$\min_{x_r} \|g(x_r) - g(x)\|_2 \qquad (۲۸)$$

در رابطهٔ (۲۸) منظور از $\boldsymbol{g}$، یک بازنمایی از ورودی و منظور از $\boldsymbol{x_r}$، ویژگی‌های مقاوم متناظر با هر نمونه‌ی $\boldsymbol{x}$ است. آن‌ها روند پیداکردن نقطهٔ بهینهٔ مسئلهٔ (۲۸) را با استفاده از کاهش گرادیان در فضای داده ورودی طبق پژوهش مادری و همکاران [۲۴] پیش برده‌اند. به بیان دیگر، آن‌ها گرادیان را در طول مراحل مختلف طبیعی می‌کردند. درواقع، آن‌ها با شناسایی ویژگی‌های مقاوم، مجموعه دادهٔ جدیدی تولید می‌کنند که شامل ویژگی‌های مقاوم است. سپس، آموزش شبکه روی مجموعه دادهٔ مقاوم صورت می‌پذیرد. آن‌ها بیان کردند که با استفاده از این نظر می‌توان صحت دسته‌بند را هم روی مجموعه داده استاندارد و هم روی نمونه‌های خصمانه افزایش داد.

نقدهای متنوعی بر پژوهش آن‌ها وجود دارد. این پژوهش تأکید زیادی بر وجود دسته‌بند مفید دارد اما ماهیت وجودی چنین دسته‌بند مفیدی ابهام دارد و معیار مشخصی برای سنجش دسته‌بند مدنظر پژوهش آن‌ها ارائه نشده‌است. این ضعف در پژوهش‌هایی نظیر **MagNet** [۵۰] نیز بیان شده که توسط طیف مختلفی از پژوهشگران نقد شده‌است.

تغییر الگو

به‌منظور دفاع در مقابل حملات خصمانه می‌توان تغییراتی در الگوی یادگیری به‌وجودآورد که براساس آن‌ها به‌صورت خودکار، دفاع انجام شود. این تغییرات می‌تواند در معماری، یا در تابع بهینه‌سازی الگو پیاده شود. برخی از تغییرات یادشده عبارت‌اند از: اصلاح عبارت منظم‌ساز[1]، خالص‌سازی دفاعی[2]، فشردن ویژگی[3]، دفاع پوشش[4] و شبکه انقباضی عمیق [5].

عبارت منظم‌ساز: عبارت منظم‌ساز اغلب با عنوان عبارت‌های جریمه[6] در تابع هزینه ظاهر می‌شوند. این عبارات قابلیت کنترل بخش‌های دیگری از هدف مدنظر را به حل مسئله بهینه‌سازی اضافه می‌کنند. بیگیو[7] و همکاران در [۵۱] تلاش کردند تا با استفاده از اضافه‌کردن عبارت منظم‌ساز، ماتریس کرنل ایجادشده در مسئله بهینه‌سازی **SVM**[8] را نسبت‌به نمونهٔ خصمانه بهینه کنند.

خالص‌سازی دفاعی: پپرنو و همکاران در [۲۸] روشی جهت دفاع در مقابل نمونه‌های خصمانه، مبتنی‌بر آموزش خالص‌سازی دانش[9] [۵۲] ارائه کردند. نظر کلی رویکرد خالص‌سازی دانش در (**شکل- ۱۰**)- **مثال انتقال دانش از الگوی معلم به الگوی دانش‌آموز در نظر پژوهش خالص‌سازی دانش [۵۲]** آورده شده‌است. همان‌طورکه می‌دانیم، برچسب تخمینی شبکه‌های عصبی در حالت عادی با دریافت لاجیت و ارسال آن به لایهٔ بیشینهٔ نرم تولید می‌شود. فرآیند خالص‌سازی دانش توسط شاخصی به نام دما (که با $\mathbf{T}$ نشان می‌دهند) انجام می‌شود. در حالت عادی مقدار این شاخص یک ($\boldsymbol{T = 1}$)، در نظر گرفته‌می‌شود. هرچه میزان $\mathbf{T}$ بیشتر باشد، خروجی توزیع نرم‌تری بین رده‌ها خواهدداشت.

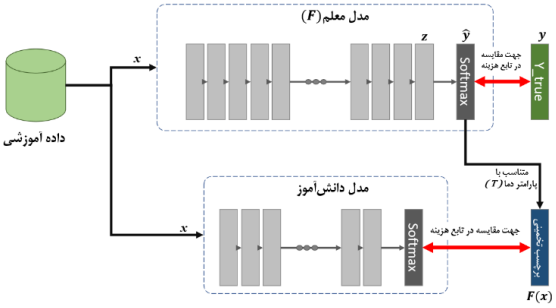
همان‌طورکه در رابطهٔ (۲۹) بیان شد، در یادگیری خالص‌سازی دانش، از شاخص نرم‌کننده‌ای به نام دما ($\boldsymbol{T}$) استفاده می‌شود. حداقل مقدار شاخص $\boldsymbol{T}$ مقدار یک است و هرچه این میزان افزایش یابد، طبق رابطهٔ (۲۹)، موجب نرم‌شدن خروجی شبکه می‌شود. بنابراین، خروجی

$$O_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_i/T)} \qquad (۲۹)$$

شبکه‌ای که در دمای بالا آموزش دیده‌باشد، نرم‌تر و متعادل‌تر است.

[1] Regularization term
[2] Defensive distillation
[3] Feature squeezing
[4] Mask defense
[5] Deep Contractive Network (DCN)
[6] Penalty terms
[7] Biggio
[8] Supported Vector Machine
[9] Knowledge distillation

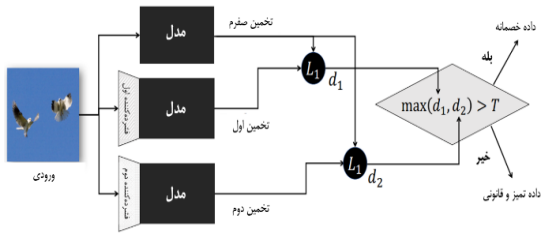
[DOI: 10.61186/jsdp.20.2.113]



(شکل- ۱۰)- مثال انتقال دانش از الگوی معلم به الگوی دانش‌آموز در نظر پژوهش خالص‌سازی دانش [۵۲]

(Figure 11) Example of knowledge transferring from teacher model to student model in the idea of knowledge distillation research [52]

روش آن‌ها از دسته روش‌های جعبه‌سفید است. نخست، شبکهٔ اولیه (شبکه معلم) با دمای مشخص $\mathbf{T}$ در بیشینهٔ نرم آموزش می‌بینند. سپس، با دادن مجموعه دادگان آموزشی $\boldsymbol{X}$ به شبکه اولیه $\mathbf{F}$ و دریافت خروجی آن، بردار احتمال شبکه اولیه به‌دست‌می‌آید. سپس، از آن بردار احتمال به‌جای برچسب مجموعه دادگان اولیه جهت آموزش همان ساختار شبکه در دمای مشخص $\boldsymbol{T}$ استفاده می‌شود. این امر موجب می‌شود تا دانش کسب‌شده از احتمال رده‌های مختلف به ازای ویژگی‌های داده ورودی بهتر بازنمایی شود. در نتیجه، به‌جای استفاده از بردارهای تک‌نقطه[۱]، که هیچ ارتباط مفهومی را بین رده‌های مختلف نشان نمی‌دهد، از بردارهای مفهوم‌دار پیش‌بینی شبکه $\boldsymbol{F}$ استفاده می‌شود. گفتنی است که موثربودن این روش بسته به شرایط خاص آموزش آن دارد. درواقع، اگر متخاصم بداند که در فرآیند آموزشی شبکه از فرآیند خالص‌سازی استفاده شده‌است، می‌تواند با رویکرد دیگری به آن حمله کند که در پژوهش کارلینی و وگنر [۵۳] به آن اشاره شده‌است.



(شکل -۱۱)- چارچوب تشخیص نمونهٔ خصمانه روش فشردگی ویژگی شو و همکاران [۵۴]

(Figure 12) The framework for detecting adversarial example of the compactness method of Xu et al. [54]

فشردن ویژگی: شو[۲] و همکاران در [۵۴] فضای قابل جستجوی متخاصم را به نحوی کاهش داده‌اند. همان‌طورکه در (شکل -۱۱) نمایان شده، آن‌ها در رویکرد پیشنهادی خود با مقایسهٔ خروجی الگو به‌ازای دادهٔ ورودی و دادهٔ فشرده‌شده با یک حد آستانه مشخص، بیان کردند که نمونهٔ خصمانه از نمونه تمیز قابل‌شناسایی است. آن‌ها بیان کردند که فشرده‌سازی ویژگی، قابلیت تعمیم و استفاده در سایر حوزه‌ها را دارد، ولی چون اغلب رویکردهای حمله در فضای دسته‌بندی استفاده قرار می‌شوند، آن‌ها نیز بر دسته‌بندی تمرکز کردند.

آن‌ها دو روش ساده برای فشرده‌کردن ویژگی در دادهٔ تصویری پیشنهاد کردند. کاهش عمق رنگی هر پیکسل در تصویر و استفاده از نرم‌کنندهٔ مکانی برای کاهش تفاوت بین پیکسل‌های مستقل، از جمله این پیشنهادها هستند. آن‌ها با انجام آزمایش‌های مختلف نشان دادند که چارچوب فشرده‌کردن ویژگی، عامل مؤثری در افزایش مقاومت الگوست. همچنین، نکتهٔ دیگر روش آن‌ها تأکید روی مراحل پیش‌پردازش به‌جای روش‌های تغییر الگوست. درواقع، آن‌ها در روش دفاع پیشنهادی، تغییری در الگو ایجاد نکرده و تنها روی ورودی خود پیش‌پردازش‌هایی را لحاظ کردند.

دفاع پوشش: گائو[۳] و همکاران در [۵۵] سعی کردند روشی جهت شناسایی ویژگی‌های مؤثّر جهت مقاوم‌کردن الگو نسبت‌به حملات ارائه کنند. آن‌ها با اضافه‌کردن یک لایهٔ پوشش قبل از لایه‌های تمام‌متصل، ویژگی‌های مؤثر را برای مرحله دسته‌بندی انتخاب و مبتنی‌بر میزان فاصله خروجی بردار ویژگی هرلایه در زمان ورود داده تمیز و داده خصمانه، اقدام به پوشش ویژگی‌های مدنظر می‌کنند.

شبکه عمیق انقباضی: گو[۴] و همکاران در [۵۶] تلاش‌های متعددی در بررسی میزان مقاوم‌بودن انواع شبکه‌های خودرمزگذار انجام دادند. آن‌ها با روش‌های متعددی نظیر افزودن نوفه به دادهٔ خصمانه و استفاده از خودرمزگذار حذف‌کنندهٔ نوفه[۵] تلاش کردند تأثیر خصمانه را کاهش دهند. ولی متوجه شدند که با این نوع آموزش، دوباره شبکه قابلیت نفوذ خواهدداشت. در نهایت، آن‌ها استفاده از خودرمزگذار قراردادی را پیشنهاد دادند. در خودرمزگذار قراردادی عبارت منظم‌ساز به تابع هزینه عادی شبکه خودرمزگذار اضافه می‌شود. رویکرد پیشنهادی آن‌ها ساختار انتهابه‌انتها دارد و رابطهٔ تابع هزینه آن به‌صورت رابطهٔ (۳۰) است:

$$J(\theta) = \sum_{i=1}^{m} \left(L(x^{(i)}, y^{(i)}) + \lambda \left\| \nabla_{x^{(i)}} h(x^{(i)}) \right\|_F^2 \right) \qquad (۳۰)$$

[۱] تنها یکی از عناصر آن بردار مقدار یک است.

[۲] Xu

[۳] Gao

[۴] Gu

[۵] Denoising Auto Encoder (DAE)

[DOI: 10.61186/jsdp.20.2.113]

که در آن منظور از m تعداد نمونه‌های آموزشی، L تابع هزینه عادی شبکه خودرمزگذار، λ شاخص عبارت منظم‌ساز و $\|\cdot\|_F$ نُرم فروبنیوس[1] است. هدف کلی تابع هزینه رابطهٔ (۳۰) ذخیره‌کردن اطلاعات مفید از دادهٔ ورودی و کاهش توجه به ویژگی‌های مختل‌شده است. آن‌ها تلاش کردند تا میزان اطلاعات یا ویژگی‌های موجود در لایه‌های مخفی را بررسی کنند و با استفاده از تقریب‌های متنوع، تابع هزینه مناسبی را در رابطهٔ (۳۱) ارائه دهند. گفتنی است آن‌ها، همچنین، اشاره کردند که تابع هزینه رابطهٔ (۳۱) بهینگی سراسری را تضمین نمی‌کند، بلکه موجب محدودکردن ظرفیت شبکه‌عصبی می‌شود.

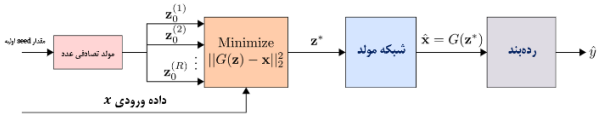
$$J(\theta)=\sum_{i=1}^{m}\left(L(t^{(i)},y^{(i)})+\sum_{j=1}^{H+1}\lambda_j\left\|\frac{\partial h_j^{(i)}}{\partial h_{j-1}^{(i)}}\right\|_2\right) \qquad (۳۱)$$

همچنین، آن‌ها بیان کردند که با افزودن نوفهٔ گوسی $\mathcal{N}(\mathbf{0},\sigma^2\boldsymbol{I})$ به ورودی و تخصیص واریانس $\boldsymbol{\sigma}\to\mathbf{0}$ به پاسخ نزدیک می‌شود.

استفاده از ابزارهای کمکی

بعضی از روش‌های مبتنی‌بر استفاده از ابزارهای جانبی عبارتند از: دفاع مبتنی‌بر شبکهٔ تقابلی مولد، شبکه مگ‌نت[2] و حذف‌کنندهٔ نوفهٔ هدایت‌شده بازنمایی سطح‌بالا[3]. در ادامه، این روش‌ها معرفی می‌شوند.

دفاع مبتنی‌بر شبکه تقابلی مولد: *سمنقویی* و همکاران در [۵۷] سازوکاری به نام **DefenseGAN** جهت دفاع نسبت‌به حملات جعبه‌سیاه، جعبه‌سفید و کاهش تأثیرگذاری اختلال خصمانه بر الگو ارائه کردند. روش آن‌ها مبتنی‌بر شبکه‌های تقابلی مولد [۵۸] است.



(شکل ۱۲) معماری شبکه تقابلی مولد استفادهشده در نظر پژوهش سمنقویی و همکاران [۵۷]

(Figure 13) The architecture of generative adversarial network used in Samangouei et al. [57]

همان‌طورکه در (شکل ۱۲) بیان شده، نظر اصلی این است که دادهٔ ورودی به فضای یادگرفته‌شدهٔ شبکهٔ مولد نگاشت داده‌می‌شود. درواقع، نزدیک‌ترین دادهٔ قابل تولید توسط مولد به عنوان دادهٔ مفید به شبکه اصلی داده می‌شود. مشکل اصلی این روش این است که نخست باید یک شبکه تقابلی مولد آموزش‌دیده موجود باشد، سپس، از آن استفاده شود. از آنجا که آموزش چنین شبکه‌ای چالش‌های خاص خود را به دنبال دارد، ترکیب آن با شبکه دیگر نیز کمی چالش‌برانگیز است. همچنین، در مرحلهٔ اول نیاز است تا بهترین $\mathbf{z}$ ممکن برای مرحله بعد انتخاب شود. زمان‌بربودن مرحلهٔ شناسایی $\mathbf{z}$ بهینه نیز از معضلات دیگر این روش به شمار می‌رود. همچنین، درصورتی‌که به شبکهٔ تقابلی مولد از قبل حمله شده باشد، این بحث به‌کلی منتفی می‌شود. بنابراین، پژوهش آن‌ها در شرایط بسیار خاصی قابلیت استفاده و اجرا دارد.

روش آن‌ها نیازمند انتخاب بهترین $\mathbf{z}$ ($\mathbf{z}^*$) و بررسی آن با حد آستانهٔ $\boldsymbol{\rho}$ موجود در رابطهٔ (۳۲) انجام می‌پذیرد. پیداکردن مقدار آستانهٔ بهینه نیز از چالش‌های دیگر پژوهش آن‌هاست.

دارای‌اختلال‌خصمانه

$$\|G(z^*)-x\|_2^2 \quad \lesseqgtr \quad \rho \qquad (۳۲)$$

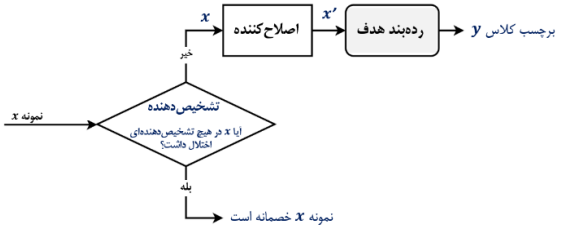
داده‌تمیز

شبکه مگ‌نت: مِنگ[4] و چِن[5] در [۵۰] چارچوبی جهت حفاظت از شبکه‌های عصبی نسبت‌به نمونه‌های خصمانه ارائه کردند. طبق (شکل ۱۳) در نظر پیشنهادی آن‌ها چند شبکهٔ تشخیص‌دهنده[6] و یک شبکهٔ اصلاح‌کننده[7] وجود دارد. شبکه‌های تشخیص‌دهنده اختلاف نمونه‌های عادی و خصمانه را مقایسه می‌کنند. شبکهٔ اصلاح‌کننده نیز داده‌های خصمانه را به منیفولد نمونه‌های عادی بازگشت می‌دهد.

مشکل اصلی نظر مگنت این است که همهٔ توانمندی آن وابسته به صحت پاسخ شبکه(های) تشخیص‌دهنده است. بنابراین، اگر به این شبکه‌ها به‌طور مجزّا حمله شود، منجر به منتفی‌شدن کل ایده و شکست آن می‌شود. بنابراین، آن‌ها پیشنهاد کردند که با تعبیهٔ چندین شبکهٔ تشخیص‌دهنده و انتخاب تصادفی هریک از آن شبکه‌ها، تاحدودی از این نوع حملات جلوگیری شود؛ ولی بعضاً در حملات جعبه‌سفید که همه شاخص‌های الگو در اختیار متخاصم است، این روش نمی‌تواند گزینهٔ چندان مناسبی جهت دفاع باشد.

[1] Frobenius
[2] MagNet
[3] High-level representation guided denoiser
[4] Meng
[5] Chen
[6] Detector
[7] Reformer

[DOI: 10.61186/jsdp.20.2.113]



(شکل ۱۳) نمایی از ایدهٔ پژوهش مگنت [۵۰]

(Figure 14) A view of MagNet research idea [50]

حذف‌کننده نوفهٔ هدایت‌شده بازنمایی سطح‌بالا:

اغلب حذف‌کننده‌های نوفهٔ استاندارد، از مشکل تشدید خطا رنج می‌برند (نظیر آن‌هایی که در سطح پیکسل کار می‌کنند). به عبارت دیگر، نوفهٔ باقیماندهٔ[1] کوچک، به‌تدریج تقویت شده و موجب خطای دسته‌بندی می‌شود. *لیائو*[2] و همکاران در [۵۹] تلاش کردند که این مشکل را با ارائهٔ تابع هزینهٔ جدید با ترکیب اختلاف خروجی الگو روی

$$D(T(x),T(\hat{x})) \ll \varepsilon \qquad (۳۵)$$

تصویر تمیز و تصویر مختل‌شده برطرف کنند. مطابق رابطهٔ (۳۳)، آن‌ها با استخراج نرم اول خروجی بازنمایی لایه $\boldsymbol{l}$ ام، برای نمونهٔ داده تمیز و مختل‌شده به سیگنالی دست‌پیدا کرده‌اند که از آن برای ادامه مسیر استفاده کردند.

$$L = \|f_l(\hat{x}) - f_l(x)\| \qquad (۳۳)$$

در پژوهش *یوساتو*[3] و همکاران [۶۰]، این روش ارزیابی و بیان شد که رویکرد مناسبی برای همه حالت‌ها نیست و این روش با دقت پایین‌تری در ارزیابی روی دادگان **ImageNet** شکست خورد.

رمزگذاری تنک کانولوشنی طبقه‌بندی‌شده:

سان[4] و همکاران در [۶۱] تلاش کردند تا با نگاشت داده‌ها به فضای ویژگی با عمق بالا و تبیین تابع هزینه مناسب، بر نمونه‌های خصمانه غالب شوند. آن‌ها مبتنی‌بر رمزگذاری تنک کانولوشنال، فضایی به‌طور تقریبی طبیعی کم‌بعدی را ایجاد و بیان کردند که نمونه‌های خصمانه به‌ندرت در آن وجود دارند. همچنین، آن‌ها با ارائهٔ لایهٔ تبدیل تنک بین تصویر ورودی و اولین لایهٔ شبکه نیز موجب بهینه‌کردن روند نگاشت پیشنهادی‌شان شدند. آن‌ها بیان کردند که در بازنمایی بالاتر، نمونه داده‌های مربوط به یک رده در فضای نزدیک به هم قرار می‌گیرند. این در حالی است که همان نمونه‌ها در فضای داده یا ویژگی اولیه، فاصلهٔ زیادی‌تری نسبت‌به هم دارند. همین امر آن‌ها را مجاب کرد که استفاده از سطح بازنمایی به عنوان یکی از رویکردهای دفاع استفاده شود. آن‌ها با دریافت ورودی و عبور آن از یک خودرمزگذار حذف‌کنندهٔ نوفهٔ از پیش‌تعلیم‌دیده، خوشهٔ مربوط به دادهٔ ورودی را شناسایی می‌کنند. سپس، خوشهٔ انتخابی و نقشه‌های ویژگی تنک (همراه با فیلترهایشان) را به‌صورت فرهنگ‌لغت[5] آماده می‌کنند. آن‌ها با استفاده از این رویکرد ورودی را به فضای تصویر به‌نسبت طبیعی هدایت می‌کنند. به بیان ریاضی، آن‌ها در تلاش هستند تا فاصله بین دادهٔ طبیعی تبدیل‌یافته ($\boldsymbol{T}(\boldsymbol{x})$) و دادهٔ خصمانهٔ تبدیل‌یافته ($\boldsymbol{T}(\hat{\boldsymbol{x}})$) را برای داده‌های یکسان کمینه کنند (در حد عدد ثابت کوچک $\boldsymbol{\varepsilon}$).

در رابطهٔ (۳۵) منظور از $\boldsymbol{D}$ فاصله و $\boldsymbol{T}$ دادهٔ ورودی به فضای طبیعی داده است. آن‌ها برای فراگرفتن فرهنگ لغت مناسب در ساختار کانولوشنی، از حل مسئلهٔ بهینه‌سازی رابطهٔ (۳۴) بهره بردند:

در رابطهٔ (۳۴) منظور از $\otimes$ عملیات کانولوشن، $\boldsymbol{C}$ تعداد کانال‌های ورودی، $\boldsymbol{K}$ تعداد فیلترهای هر کانال ورودی، $\boldsymbol{f_{i,c}}$ مجموعه‌ای از فیلترها و $\mathbf{z_i}$ نقشه ویژگی هر فیلتر است. درواقع، برخلاف اغلب روش‌های مرسوم رمزگذاری تنک استاندارد که فرهنگ لغت و رمز برای کل داده فراگرفته می‌شد، تلاش کردند تا بازسازی تکه‌های تصویر، توسط رمز و فرهنگ لغت محلی انجام شود. رویکرد آن‌ها در مقابل، با روش‌های حملهٔ مرسوم نظیر **FGSM**، **BIM**، **DeepFool** و **C&W** عملکرد خوبی داشته است.

$$\begin{aligned} &\min \frac{1}{2}\sum_{c=1}^{C}\|x_c - T(x)_c\|_2^2 + \lambda\sum_{i=1}^{K}\|z_i\|_1 \\ &s.t.\ \ T(x)_c = \sum_{i=1}^{K} f_{i,c}\otimes z_i \\ &s.t.\ \ \|f_{i,c}\|_2^2 = 1 \\ &\quad 1 \le i \le K, 1 \le c \le C \end{aligned} \qquad (۳۴)$$

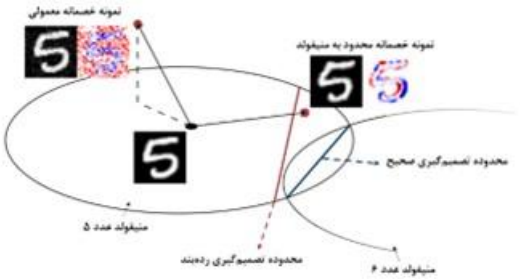
۴- بحث در مورد تحلیل درهم‌تنیدگی ویژگی‌ها و تأثیر آن در تعمیم‌پذیری و مقاوم‌سازی

در این قسمت به مرور برخی پژوهش‌ها در خصوص تحلیل ویژگی‌های فراگرفته‌شده در الگوهای یادگیری ماشین با تأکید بر مسئله تحلیل حساسیت نسبت‌به نمونه‌های خصمانه پرداخته شده‌است. *سیپراس*[6] و

[1] Residual
[2] Liao
[3] Uesato
[4] Sun
[5] Dictionary
[6] Tsipras

[DOI: 10.61186/jsdp.20.2.113]

همکاران در [۶۲] به وجود یک تنش بین تعمیم و مقاوم‌سازی الگوی یادگیری اشاره کردند. آن‌ها بیان کردند که آموزش مقاوم موجب کاهش معیار ارزیابی صحت می‌شود. بنابراین، باید برای رسیدن به پاسخ بهینه بین تعمیم‌پذیری و مقاوم‌سازی الگو مصالحه ایجاد کرد و بررسی سطحی آن‌ها موجب رسیدن به پاسخ موجهی نمی‌شود. همچنین، آن‌ها اشاره‌کردند که مقاوم‌سازی در نتیجهٔ آموزش عادی پدید نمی‌آید و نیاز است با به‌کارگیری یک رویکرد بهینه انجام شود.



(شکل ۱۴)- طرحی از نظر تحلیلی اشتوتز و همکاران [۳۸]

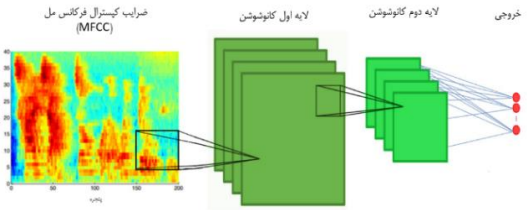
(Figure 15) Analytical view for the idea of Stutz et al. [38]

از طرفی آن‌ها بیان کردند که یادگیری ویژگی پدیدآمده توسط چارچوب شبکه تقابلی مولد می‌تواند با مقاومت‌پذیری ارتباط تنگاتنگی داشته‌باشد. *سو*[1] و همکاران در [۶۳] نیز تأکید کردند که بین صحت و مقاومت‌پذیری مصالحه‌ای وجود دارد و بالابردن میزان صحت، کاهش مقاومت‌پذیری را به‌دنبال‌دارد. آن‌ها ادعای یادشده را روی ۱۸ معماری محبوب مجموعه دادگان **ImageNet** پیاده و صحت مصالحه را تأیید کردند. *اشتوتز*[2] و همکاران در [۳۸] سعی کردند تا ارتباط بین تعمیم‌پذیری و مقاومت الگوها را بررسی کنند (**شکل ۱۴**).

آن‌ها تعریف جدیدی از نمونه‌های خصمانه بیان و نمونه‌های خصمانه را به دو دسته نمونه‌های خصمانهٔ معمولی[3] و نمونه‌های خصمانهٔ محدود به منیفولد[4] تقسیم‌بندی و بیان کردند که نمونه‌های خصمانهٔ معمولی روی منیفولد قرار ندارند و مقاوم‌سازی نسبت‌به آن‌ها مفید بوده و هیچ تداخلی با تعمیم‌پذیری ندارد. اما نمونه‌های خصمانه محدود به منیفولد موجب افزایش تعمیم‌پذیری و همچنین، افزایش صحت می‌شود. درواقع، وجود نمونه‌های خصمانهٔ محدود به منیفولد ناشی از عدم یادگیری صحیح دسته‌بند است و این امر موجب می‌شود تا محدودهٔ تصمیم‌گیری دسته‌بند از محدوده تصمیم‌گیری درست فاصله داشته‌باشد.

به بیان دیگر، آن‌ها بیان کردند که نمونه‌های خصمانهٔ محدود به منیفولد، ناشی از خطای تعمیم‌پذیری بوده‌است. تأکید پژوهش آن‌ها بر عدم تداخل بین مقاومت‌پذیری و تعمیم‌پذیری الگوست. آن‌ها تفاوت نمونه‌های خصمانهٔ معمولی از نمونه‌های خصمانهٔ محدود به منیفولد را با محاسبهٔ فاصله بین نمونه اصلی و تقریب تصویرسازی متعامد[5] نمونه بر منیفولد به‌دست‌آوردند.

باتوجه‌به‌این‌که در حمله‌های جعبه‌سیاه دسترسی به الگو امکان‌پذیر نیست، به‌طور معمول، از الگوی جایگزین برای به‌دست‌آوردن اختلال و تولید نمونهٔ خصمانه استفاده می‌شود. شباهت بازنمایی تولیدشده در الگوی اصلی با بازنمایی ویژگی در الگوی جایگزین و همچنین، قابلیت فریب آن‌ها از موضوعات چالشی این حوزهٔ پژوهشی است. گودفلو و همکاران در [۸] بیان کردند که نمونه‌های خصمانه اغلب در نواحی همجوار زیرفضای یک‌بعدی که توسط روش علامت گرادیان سریع تعریف می‌شود، رخ می‌دهند. آن‌ها نمونه‌های خصمانه را توسط روش‌های مختلف نظیر علامت گرادیان سریع تولید کردند. همچنین، ترامر و همکاران در [۶۴] به‌صورت تقریبی نشان دادند که نمونه‌های خصمانه در زیرفضای همجوار با ابعاد به‌نسبت بالا[6] قرار دارند. درواقع، آن‌ها بیان کردند که در ابعاد بالاتر بیشترِ زیرفضاهای یافت‌شده دارای اشتراک هستند و همدیگر را پوشش می‌دهند. زیرفضای مشترک موجب انتقال‌پذیری نمونهٔ خصمانه از یک الگو به الگوی دیگر می‌شود. آن‌ها بیان کردند که برای وظیفهٔ خاص، معماری‌های مختلف آموزش‌دیدهٔ شبکهٔ عصبی دارای زیرفضاهای مشترک هستند. در دیدگاه آن‌ها، به محدودهٔ تصمیم‌گیری توجه شده‌است.



(شکل ۱۵)- ساختار یک سامانهٔ پردازش گفتار مبتنی‌بر استفاده از شبکه‌های عصبی کانولوشنی

(Figure 16) Structure of a speech processing system using convolutional neural networks

کاریاپا[7] و *قریشی* در [۱۳] به مسئلهٔ کاهش انتقال‌پذیری نمونه‌های خصمانه پرداختند. آن‌ها تلاش کردند تا با ارائهٔ تابع هزینه تنظیم گرادیان[8] و استفاده از یادگیری جمعی، موجب کاهش فضای اشتراکی نمونه‌های

[1] Su
[2] Stutz
[3] Regular
[4] Constrained to the manifold

[5] Orthogonal projection
[6] در پژوهش ترامر و همکاران ذکر شده که این ابعاد حدود ۲۵ بعد است.
[7] Kariyappa
[8] Gradient Alignment Loss (GAL)

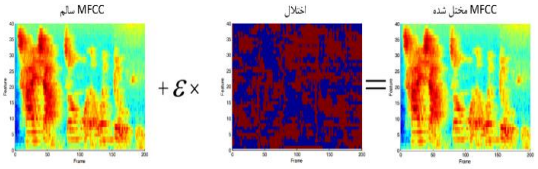
[DOI: 10.61186/jsdp.20.2.113]

خصمانه بین الگوهای مختلف یادگیری جمعی شوند. همان‌طورکه ترامر و همکاران در [۶۴] بیان کردند، زیرفضایی که نمونه‌های خصمانه تشکیل می‌دهند، دارای ابعاد مشخصی است. اگر تعداد ابعاد افزایش یابد، اشتراک زیادی در ابعاد بالا با ابعاد پایه ایجاد می‌شود. در یادگیری جمعی نیز این امر طبق بیان کاریاپاپا و قریشی در [۱۳] تأیید شده‌است. هرچه این زیرفضاهای خصمانه دارای اشتراک کمتری باشند، انتقال‌پذیری کمتری بین الگوهای مختلف ایجاد می‌شود و موجب می‌شود تا متخاصم سخت‌تر الگوی کلی را فریب دهد. از شباهت کسینوسی برای تنظیم‌کردن گرادیان استفاده می‌شود.

۵- حمله و دفاع در حوزه‌های غیر از بینایی ماشین

حوزه‌های غیراز بینایی ماشین نظیر پردازش گفتار، متن و... نیز از اختلال‌های خصمانه مصون نیستند. درواقع، افراد متخاصم در همهٔ حوزه‌ها به‌دنبال پیداکردن اختلال‌های بهینه جهت فریب الگو هستند.

در غالب پژوهش‌های حوزهٔ گفتار با استخراج ضرایب بانک فیلتر ملبرای فریم‌های متوالی، دادهٔ صوتی ورودی را به تصویر تبدیل می‌کنند ((**شکل ۱۵** و (**شکل ۱۶**)). در پژوهش *ونگ*[1] و همکاران از آن برای تشخیص کلیدواژهٔ صوتی[2] استفاده کردند. آن‌ها تلاش کردند تا همانند پژوهش‌های حوزهٔ بینایی ماشین به تولید نمونهٔ خصمانهٔ صوتی روی آورده و از این طریق ردهٔ کلیدواژه صوتی ورودی را تغییر دهند.



(شکل ۱۶)- نمایی از مختل‌کردن ضرایب کپسترال فرکانس مل با استفاده از نمونه‌های خصمانه [۶۵]

(Figure 17) Perturbing Mel-frequency cepstral coefficients using adversarial examples [65]

رویکرد حملهٔ آن‌ها در راستای استفاده از شبکه رقابتی مولد تعریف می‌شود. در مرحله دفاع، آن‌ها با استفاده از روش آموزش خصمانه اقدام به آموزش مجدد روی مجموعه داده ترکیبی سالم و خصمانه کردند [۶۵].

گفتنی است که در خلال روند پیشروی از روش‌های هم‌ترازی تحمیلی[3] و محاسبهٔ حد آستانهٔ شنوایی نیز در روش‌هایی نظیر روش *شانهر*[4] و همکاران در [۶۶] استفاده شده‌است تا صوت خروجی با عامل انسانی همچنان قابل تشخیص بوده، و تغییرات محسوس نباشد. همین روند در حوزهٔ پردازش متن به شکل مشابهی وجود دارد. در این حوزه با تغییر حروف، یا کم و زیاد کردن علائم نگارشی اختلال‌های موردنظر پدید می‌آید.

در (**شکل ۱۷**) نمونه‌ای از اختلال‌های خصمانه در حوزهٔ پردازش متن آورده شده‌است. مشاهده می‌شود که با بزرگ‌شدن یک حرف، دستهٔ جملهٔ موردنظر تغییر کرده‌است. (**شکل۲۳**) تقسیم‌بندی انواع رویکردهای دفاع بررسی‌شده را نشان می‌دهد.

۶- آزمایش‌ها

در ادامه با انجام برخی آزمایش‌ها، به مقایسه و ارزیابی برخی روش‌های حمله و دفاع مورد ارزیابی پرداخته‌ایم. از دو مجموعهٔ دادگان در این آزمایش‌ها استفاده شده‌است.

South Africa's historic Soweto township marks its 100th birthday on Tuesday in a mood of optimism.
57% **World**

South Africa's historic Soweto township marks its 100th birthday on Tuesday in a moo**P** of optimism.
95% **Sci/Tech**

Chancellor Gordon Brown has sought to quell speculation over who should run the Labour Party and turned the attack on the opposition Conservatives.
75% **World**

Chancellor Gordon Brown has sought to quell speculation over who should run the Labour Party and turned the attack on the o**B**position Conservatives.
94% **Business**

(شکل ۱۷)-نمایی از مثال نمونهٔ خصمانه در حوزه متن و کار دسته‌بندی متن [۶۷]

(Figure 18) A scheme of an adversarial example in the text domain and text classification task [67]

مجموعهٔ دادگان

در این پژوهش از مجموعهٔ دادگان **MNIST** [۶۸] و **CIFAR10** [۶۹] استفاده شده‌است. مجموعه داده **MNIST** متشکل از ۶۰ هزار نمونهٔ اعداد دستنویس انگلیسی و به‌صورت سطوح خاکستری (تک‌کانال رنگی) است. مجموعه داده **CIFAR10** از ۶۰ هزار نمونه تصویر رنگی (سه کانال رنگی) در ده ردهٔ مختلف تشکیل شده‌است. تقسیم‌بندی این دادگان‌ها به دو بخش آموزشی و آزمون در این پژوهش، طبق

[1] Wang
[2] Keyword spotting
[3] Force alignment
[4] Schönherr

[DOI: 10.61186/jsdp.20.2.113]

تقسیم‌بندی انجام‌شده توسط فراهم‌کنندگان هریک از این دادگان‌هاست.

معماری

در این پژوهش از معماری **WRN-28** مطرح‌شده توسط *زاگوروکو*[1] و *کوموداکیس*[2] استفاده شده‌است [۷۰]. این معماری دارای ۲۸ بلوک کانولوشنی باقیمانده[3] با خاصیت **wide-dropout** با شامل ۳۶/۵ میلیون شاخص بوده و با ارزیابی انجام‌شده با تعداد ۲۰۰ تکرار، روی مجموعه داده **CIFAR10** دارای صحت ۹۶/۰۹ درصد است. برای مجموعه داده **MNIST** از معماری معرفی‌شده توسط *سیلوا*[4] در [۷۱] استفاده شده که با تعداد ۲۰۰ تکرار روی مجموعه داده **MNIST** دارای صحت ۹۹/۲۱ درصد است.

مشخصات سیستم پردازشی

برای انجام مجموعه آزمایش‌ها از یک سرور پردازشی سیستم عامل **Ubuntu 18.04**، پردازنده **AMD FX(tm)-8350**، پردازندهٔ گرافیکی **GeForce GTX 1080 Ti** و حافظهٔ دسترسی تصادفی ۳۲ گیگابایت، همچنین، از زبان پایتون[5] نسخه ۳/۶ برای رمزنویسی و کتابخانهٔ پایتورچ[6] [۷۲] نسخه ۱/۶ نیز برای نوشتن بستر آزمایش‌ها استفاده شده‌است.

(جدول ۲) صحت الگوی WRN-28 آموزش‌دیده روی مجموعه داده MNIST با نمونه‌های خصمانهٔ تولیدشده از روش‌های حملهٔ مختلف تحت نرم l_∞

(Table 2) Accuracy of the WRN-28 trained model for the MNIST dataset with adversarial example generated by different attack methods under l_∞ norm

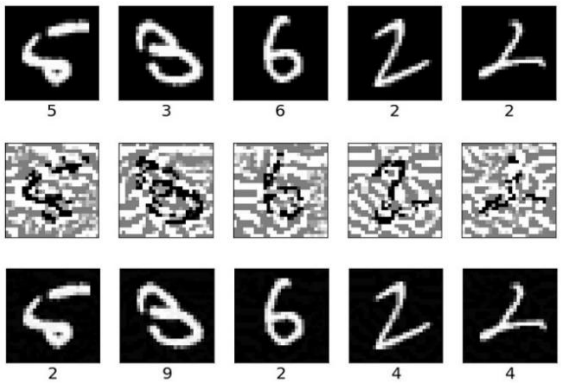
روش حمله \ ε	صحت الگوی مبنا	۰/۰۱	۰/۰۳	۰/۱	۰/۲	۰/۳
FGSM		۹۸/۴۴	۹۶/۵۱	۷۳/۵۶	۲۳/۵۹	۱۳/۳۲
PGD		**۹۸/۳۹**	**۹۲/۵۲**	**۸/۲۵**	**۰/۲۴**	**۰/۱۰**
FFGSM		۹۸/۴۲	۹۶/۴۲	۸۴/۲۸	۳۸/۷۴	۱۲/۰۳

مقایسهٔ برخی از پرکاربردترین حملات خصمانه

در ادامه به بررسی چند روش پرکاربرد حمله روی دو معماری یادشده به‌همراه مجموعه دادگان مرتبط با آن‌ها خواهیم‌پرداخت. همه آزمایش‌های انجام‌شده در معماری‌های مرتبط در شرایط یکسان انجام شده‌اند. در (**جدول ۲** نتایج صحت حملات به الگوی **WRN-28** آموزش‌دیده آورده شده‌است. در ستون سمت چپ این جدول اسامی حملات قید شده‌است. در (**جدول ۴** صحت الگوی آموزش‌دیده روی مجموعه داده **MNIST** نسبت‌به نمونه‌های خصمانه تولیدشده از روش‌های حمله **FGSM**، **PGD** و **FFGSM** با میزان اختلال‌های مشخص ۰/۰۱ الی ۰/۳ آورده شده‌است. مشخص است که با افزایش میزان اختلال، صحت الگوی اولیه کاهش می‌یابد. همچنین، مشخص است که حمله **PGD** با تعداد ۵۰ مرحله و میزان شاخص آلفای ۰/۰۷ بهتر از روش‌های دیگر تحت l_∞ توانسته صحت الگو را کاهش دهد. درواقع، روش **PGD** به خاطر مرحله‌ای‌بودن آن و همچنین، تصویرکردن روی محدوده l_∞ بهتر به مرز دسته‌بند نزدیک شده و به سمت دیگر مرز دسته‌بند هدایت می‌شود. در (**شکل ۱۸**) نمایی از نمونه‌های سالم، اختلال و نمونه‌های خصمانهٔ تولیدشده از روش **PGD** با تعداد تکرار ۵۰ مرحله روی مجموعه داده **MNIST** نشان داده شده‌است. در زیر هر تصویر، پاسخ‌های الگوی آموزش‌دیده نشان داده شده‌اند.

همان‌طورکه در سطر اول (**شکل ۱۸**) مشخص شده، نمونه‌های سالم به‌درستی توسط الگو مشخص شده‌است. اما با اضافه‌شدن اختلال حاصل از روش حمله **PGD-50** (که در سطر دوم نمایی از آن با کمی افزایش مقادیر برای مشخص‌شدن اختلال‌ها مشخص شده) به نمونه‌های خصمانهٔ سطر سوم دست پیدا می‌کنیم. همان‌طورکه مشخص است، همه برچسب‌های نمونه‌های خصمانه توسط مدل به اشتباه شناسایی شده‌است. چنانچه آزمایش‌های انجام‌شده را روی فضای رنگی مجموعه داده **CIFAR10** با معماری ذکرشده در بخش ۰ آزمایش کنیم، نتایج صحت حملات مبتنی‌بر محدوده l_∞ نظیر **FGSM**، **PGD** و **FFGSM** با میزان اختلال‌های ۰/۰۱ الی ۰/۳ و به الگوی آموزش‌دیده به‌صورت

(**جدول ۳** خلاصه می‌شود. نتایج حملات مبتنی‌بر محدوده به l_2 پس از این بخش بررسی شده‌است.



(شکل ۱۸)- نمایی از نمونه‌های سالم (ردیف بالا)، اختلال‌های تولیدشده توسط روش حمله PGD-50 (ردیف وسط) و

[1] Zagoruyko
[2] Komodakis
[3] Residual
[4] Silva
[5] Python
[6] PyTorch

[DOI: 10.61186/jsdp.20.2.113]

نمونه‌های خصمانهٔ حاصل شده از آن حمله (ردیف پایین) برای الگوی آموزش‌دیده روی مجموعه داده MNIST

(Figure 19) A view of clean samples (top row), generated perturbation with PGD-50 attack method (middle row) and adversarial examples obtained from that attack (bottom row) for the model trained on MNIST dataset

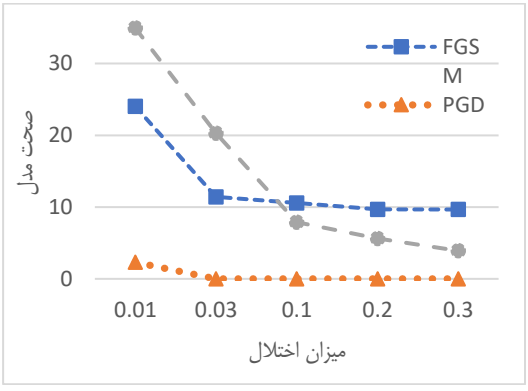
(جدول ۳)- صحت الگوی آموزش‌دیده روی مجموعه داده CIFAR10 با نمونه‌های خصمانه تولیدشده از روش‌های حملهٔ مختلف تحت نرم l_∞

(Table 3) Accuracy of the trained model for the CIFAR10 dataset with adversarial example generated by different attack methods under l_∞ norm

	صحت الگوی مبنا				
روش حمله\ε	۰/۰۱	۰/۰۳	۰/۱	۰/۲	۰/۳
FGSM	۲۴/۰۳	۱۱/۴۴	۱۰/۵۹	۹/۷۰	۹/۶۶
PGD	**۲/۳۱**	۰/۰۰	۰/۰۰	۰/۰۰	۰/۰۰
FFGSM	۳۴/۹۴	۲۰/۳۰	۷/۹۰	۵/۶۳	۳/۹۱

همان‌طورکه در (**جدول ۳** آورده‌شده، صحت الگوی آموزش‌دیده روی مجموعه داده **CIFAR10** نسبت‌به حملات مختلف محصورشده به نرم l_∞ نظیر **FGSM**، **PGD** و **FFGSM** با میزان اختلال‌های ۰/۰۱ الی ۰/۳ ارزیابی شده‌است.

نتایج نشان می‌دهد که روش **PGD** یا **PGD-50** با تعداد ۵۰ تکرار از روش‌های دیگر قدرتمندتر است. یکی از دلایل اصلی آن، رویکرد چندمرحله‌ای‌بودن آن است و حتی در میزان اختلال‌های بزرگ‌تر، دوباره روش **PGD** صحت الگو را بیشتر کاهش می‌دهد (شکل ۲۰).



(شکل ۱۹) مقایسهٔ قدرت حملات محصورشده[1] مبتنی‌بر نرم l_∞ روی الگوی آموزش‌دیده با مجموعه داده MNIST

(Figure 20) Comparison of the strength of l_∞ bounded attacks for the model trained using MNIST dataset

روش‌های مبتنی‌بر l_2 نظیر **CW** و همچنین، **DeepFool** باتوجه به ماهیت رویکرد حمله آن‌ها به‌طور مجزّا در (**جدول ۴** ارزیابی شده‌اند. در این روش‌ها تکرارپذیربودن آن‌ها از یک طرف و زمان اجرای حمله آن‌ها از طرف دیگر نیز تأثیرگذار بوده و این مدت‌زمان در روش **DeepFool** به‌خاطر ماهیت سخت‌گیری میزان اختلال به‌مراتب بالاتر است. درواقع، روش **DeepFool** نسبت‌به روش‌های دیگر اختلال بسیار کوچکتری را دارد. بحث زمان‌بربودن محاسبات آن نیز توسط موسوی دزفولی و همکاران در [۱۵] مطرح شده‌است. آن‌ها بیان کردند که رویکردهای دیگر نسبت‌به **DeepFool** در مدت‌زمان اجرا حدود ۵ برابر سریع‌تر هستند. در آزمایش‌های انجام‌شده مطابق (جدول ۴ نیز این بیان تأییدشده و روش **DeepFool** نسبت‌به حمله **CW** به‌مراتب در اجرا طولانی‌تر است.

(جدول ۴)- صحت الگوی آموزش‌دیده روی مجموعه دادگان MNIST و CIFAR10 و زمان اجرای روش‌های حملهٔ مختلف تحت محدودهٔ محصورشده به l_2

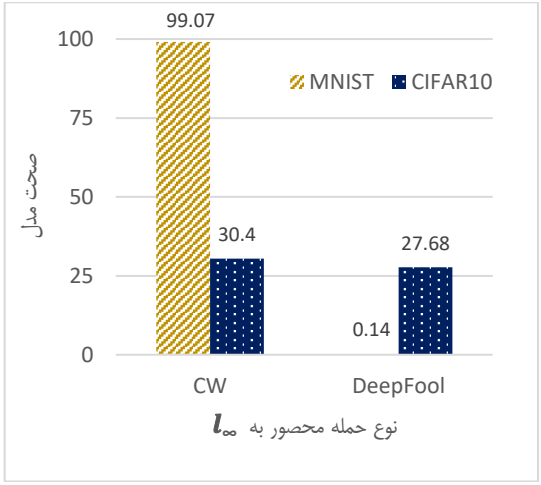
(Table 4) Accuracy of the model trained using MNIST and CIFAR10 datasets and executive time for different l_2 bounded attacks

	MNIST		CIFAR10	
روش‌حمله	صحت	زمان (S)	صحت	زمان (S)
CW	۹۹/۰۷	۶/۴۷	**۰/۱۴**	۸۰۵/۷۸
DeepFool	**۳۰/۴۰**	**۱۵۳۷۵/۴**	۲۷/۶۸	**۴۶۳۴۳/۵**

باتوجه به توضیحاتی که بیان شد، و همان‌طورکه در (**جدول ۴** آورده‌شده، مدت زمان اجرای حمله **DeepFool** روی کل مجموعه داده آزمون **MNIST** در حدود ۱۵۳۷۵.۴ ثانیه است. این میزان به‌ازای هر نمونه حدود ۱/۵ ثانیه طول می‌کشد و نسبت‌به سایر رویکردهای حمله نظیر حملهٔ **CW** به‌مراتب بیشتر است. البته قدرت حملهٔ الگوریتم **DeepFool** به‌مراتب از **CW** بیشتر است. همان‌طورکه مشاهده می‌شود، در مقایسه با حملهٔ **CW** با ۵۰ تکرار، روش **DeepFool** توانایی کاهش صحت بیشتری نسبت‌به **CW** روی مجموعه داده **MNIST** داشته‌است.

[1] Bounded

[DOI: 10.61186/jsdp.20.2.113]

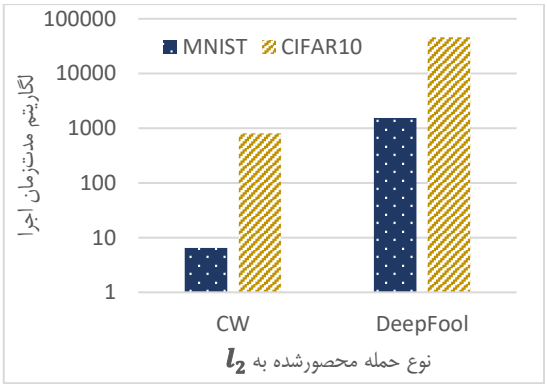


(شکل ۲۰)- مقایسه قدرت حملات محصورشده مبتنی‌بر نرم l_2 روی الگوی آموزش‌دیده با مجموعه دادهٔ MNIST و CIFAR10

(Figure 21) Comparison of the strength of l_2 bounded attacks for the model trained using MNIST and CIFAR10 datasets

این نتایج روی مجموعه‌داده **CIFAR10** نیز ارزیابی و نتایج آن در ستون دوم (**جدول ۴** درج شده‌است. همان‌طورکه مشخص‌شده، مدت زمان اجرای حملات روی مجموعه‌داده **CIFAR10** به‌واسطهٔ افزایش اندازه از ۲۸×۲۸ پیکسل در نمونه داده **MNIST** به ۳۲×۳۲ پیکسل در نمونه داده **CIFAR10** طولانی‌تر شده‌است. همچنین، مشخص است که مدت‌زمان اجرای حمله **DeepFool** نسبت‌به حملهٔ **CW**، بسیار طولانی‌مدت بوده‌است (در حدود ۴۶۱۴۳/۵ ثانیه).

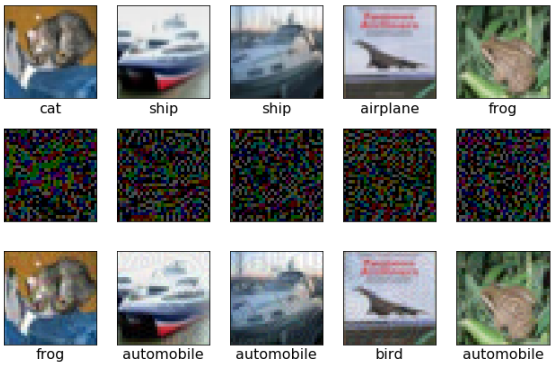
(**شکل ۲۱**) مقایسهٔ مدت‌زمان اجرای روش **DeepFool** و **CW** روی دو الگوی آموزش‌دیده روی **MNIST** و **CIFAR10** را نشان می‌دهد. باتوجه‌به زیادبودن بازّ مدت‌زمان اجرای روش **DeepFool**، میزان زمان به لگاریتم مبنای ۱۰ تبدیل شده‌است.



(شکل ۲۱)- مقایسه لگاریتم مدت‌زمان اجرای حملات محصورشده به l_2

(Figure 22) Comparison of execution time of the l_2 bounded attacks

در (شکل ۲۲) نمایی از نمونه‌های سالم، اختلال تولیدشده توسط روش حمله **PGD-50** و همچنین، نمونه‌های خصمانهٔ حاصل‌شده از آن حمله برای الگوی آموزش‌دیده روی مجموعه داده **CIFAR10** آورده شده‌است.



(شکل ۲۲)- نمایی از نمونه‌های سالم (ردیف بالا)، اختلال (ردیف وسط) و نمونهٔ خصمانه تولیدشده توسط روش حملهٔ PGD-50 (ردیف پایین) برای الگوی آموزش‌دیده روی مجموعه داده CIFAR10

(Figure 23)) A view of clean samples (top row), generated perturbation (middle row) and adversarial examples obtained from PGD-50 attack method (bottom row) using the model trained using CIFAR10 dataset

همان‌طورکه در سطر نخست، (**شکل ۲۲**) مشخص شده، نمونه‌های سالم به‌درستی توسط الگو شناسایی شده‌اند. با اضافه‌شدن اختلال حاصل از روش حمله **PGD-50**، به نمونه‌های خصمانه سطر سوم دست پیدا می‌کنیم. همهٔ برچسب‌های نمونه‌های خصمانه توسط الگو به‌اشتباه شناسایی شده‌است. یکی از نکات جالب این است که برای یک ردهٔ واحد، با اختلال‌های مختلف **PGD-50** روی هر داده تصویری **MNIST** از برچسب کشتی(**ship**)، نمونهٔ خصمانه متناظر با آن‌ها به ردهٔ خودرو (**automobile**) تغییر برچسب داده‌است.

مقایسهٔ برخی روش‌های پرکاربرد دفاع نسبت‌به حملات خصمانه

باتوجه‌به‌اینکه رویکرد یادگیری خصمانه به عنوان روشی پایدار و مقاوم نسبت‌به حملات خصمانه معرفی شده و نسبت‌به سایر رویکردها برتری داشته، در ادامه طی آزمایش‌هایی تأثیر استفاده از آموزش خصمانه بررسی شده‌است.

همان‌طورکه در (**جدول ۵** مشخص شده صحت الگوی پایهٔ آموزش‌دیده روی مجموعه داده **CIFAR10** نسبت‌به حملات خصمانه **FGSM**، **PGD** و **FFGSM** با شاخص میزان اختلال ۰/۰۳ نشان داده شده‌است. زمانی‌که

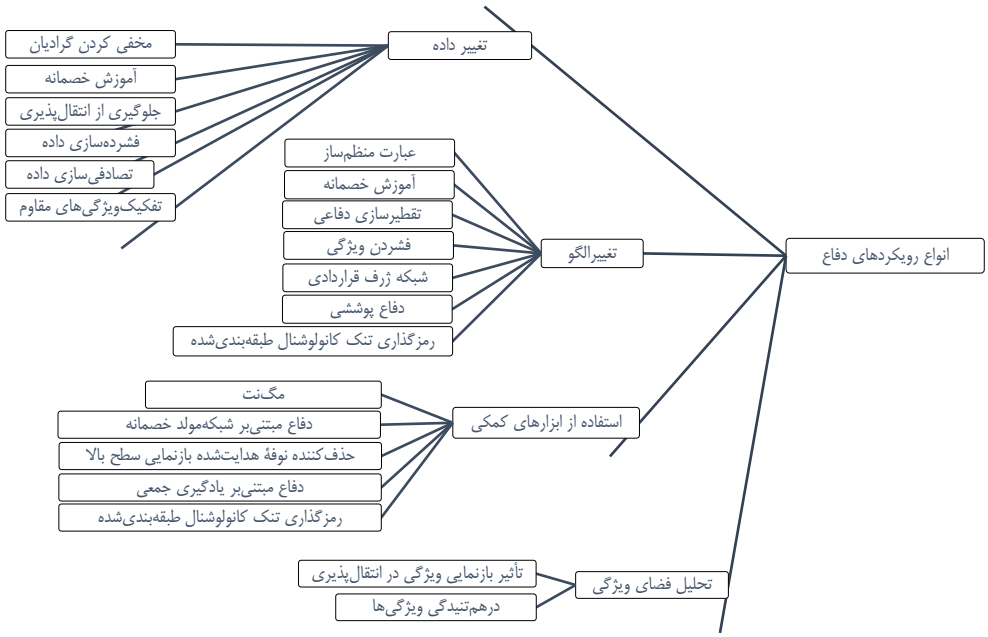
در حالت عادی شبکه با ۲۰۰ تکرار آموزش می‌بینند، صحت الگو روی نمونه‌های آزمون ۹۶/۰۹ درصد است. چنانچه همین الگو با استفاده از آموزش خصمانه و حمله **FGSM** با میزان اختلال ۰/۰۳ آموزش بیند، صحت الگو روی مجموعه داده آزمون به میزان ۸۳/۵۹ درصد می‌رسد.

(جدول ۵)- مقایسهٔ تأثیر صحت الگوی عادی و الگوی آموزش‌دیده مبتنی‌بر آموزش خصمانه با حملات مختلف روی مجموعه داده CIFAR10

(Table 5)– Comparison of the accuracy of the model trained over different attacks using CIFAR10 dataset

	صحت الگو			
حمله/ الگو	-	FGSM	PGD	FFGSM
Base	**۹۶/۰۹**	۱۱/۴۴	۰/۰۰	۲۰/۳۰
Adv- FGSM	۸۳/۵۹	**۹۹/۱۱**	۰/۰۰	۲۵/۵۴
Adv-PGD	۹۱/۴۰	۵۸/۹۰	**۴۹/۲۲**	**۸۰/۴۰**
Adv-FFGSM	۷۸/۹۰	۹۵/۸۶	۴/۴۰	۴/۴۰

همین روال آموزش خصمانه با حملات **PGD** و **FFGSM** نیز انجام شده و نتایج صحت الگوی آموزش‌دیده روی مجموعه داده آزمون به‌ترتیب ۹۱/۴۰ و ۷۸/۹ به‌دست‌آمده است. مشخص است که صحت در آموزش خصمانه به‌نسبت کاهش یافته و این امر تاحدودی پدیدهٔ منطقی و عادی است. زمانی‌که به سمت مقاوم‌پذیری حرکت می‌کنیم، لزوماً صحت الگو افزایش ندارد، بلکه کاهش هم خواهدداشت و باید نقطهٔ مصالحه هر دو معیار را شناسایی کرد. الگوی آموزش‌یافته مبتنی‌بر آموزش خصمانه با حمله **FGSM** (**adv-FGSM**)، نسبت‌به حملهٔ **FGSM** نسبت‌به بقیه و به‌ویژه حالت پایه مقاوم‌تر شده ولی نسبت‌به سایر حملات همچنان مقاومتی ندارد. از بین همه الگوهای آموزش‌یافته مبتنی‌بر آموزش خصمانه، الگوی آموزش‌دیده مبتنی‌بر حمله **PGD** با تعداد



(شکل۲۳) خلاصه‌ای از انواع رویکردهای دفاع [۶]، [۱۱] و [۳۸]

(Figure 24) A summary of different defense approaches

تکرار ۲۰۰ و تعداد مراحل ۵۰ و آلفای ۰/۰۲ و میزان اختلال ۰/۰۳ دارای بالاترین صحت هم‌زمانی در بین حملات مختلف است. درواقع، این امر حاکی از این است که حملهٔ **PGD** نسبت‌به حمله **FGSM** قوی‌تر بوده و روی نقاط بسیار حساس‌تری تمرکز می‌کند و چنانچه این نقاط حساس را در حین آموزش شبکه در نظر بگیریم و تابع هزینه را نسبت‌به آن‌ها تعریف نماییم، موجب می‌شود که الگوی غالب جوانب مختلفی از بازنمایی داده‌ها را شناسایی کند و روی نقاط حساس هم برای تشخیص برچسب و کلاس متمرکز و از این رو مقاوم‌پذیری آن نسبت‌به طیف بیشتری از حملات مقاوم شود. هدف آزمایش انجام‌شده مقایسهٔ آموزش خصمانه با حملات مختلف است و به همین جهت تنها با شاخص‌های مذکور همه مقایسه‌ها انجام شده و هدف، پیداکردن شاخص بهینه نبوده‌است. بنابراین، احتمال می‌رود که با شناسایی شاخص‌های بهینهٔ هر روش حمله، هم در مرحلهٔ آموزش

[DOI: 10.61186/jsdp.20.2.113]

خصمانه و هم در مرحلهٔ آزمون، شاهد صحت‌های بهتری باشیم. باتوجه‌به زمان‌بربودن روش آموزش خصمانه در هر تکرار از آموزش، باید به آن تکه‌داده[1] حمله شود و چنانچه حمله زمان‌بر باشد، کل آموزش زمان‌بر خواهدشد. این امر در آموزش خصمانه با حملهٔ **PGD** به‌مراتب از همهٔ حملاتِ ارزیابی‌شده در آموزش خصمانه طولانی‌تر بوده‌است. بنابراین، در روش آموزش خصمانه که به‌عنوان یکی از برترین رویکردهای دفاع شناخته می‌شود، نوع حملهٔ مورداستفاده بسیار مهم بوده و شناسایی شاخص‌های بهینه آن حمله نیز مهم است.

۶- نتیجه‌گیری

بررسی و ارزیابی رویکردهای مختلف حمله و دفاع از شبکه‌های عصبی عمیق نسبت‌به حملات خصمانه برای تولید الگویی پایدار و قابل اعتماد مهم است. تمرکز و حل چالش‌های این حوزه موجب تولید الگوی مقاوم نسبت‌به حملات خصمانه می‌شود. همین امر در صورت امکان می‌تواند با افزایش صحت و درک دقیق از سازوکار شبکه‌های عصبی عمیق همراه شود. در این پژوهش عوامل و ویژگی‌های مؤثر در آسیب‌پذیری شبکه‌های عصبی عمیق بررسی شد. به بیان دیگر، تلاش شد تا با مروری ساختارمند در حوزهٔ یادگیری ماشین خصمانه، بخشی از مهمترین مفاهیم آن بازگو و نظرات مختلف به تحلیل و بررسی گذاشته شود. پس از ارائهٔ مفاهیم کلیدی این حوزه به بررسی و نقد رویکردهای متنوع حمله به شبکه‌های عصبی عمیق پرداخته‌شد. مشخص شد که حملات تکرارشونده از حملات عادی الزاماً بهتر نیستند، ولی تعریف مسئله روش **DeepFool** در l_2 و همچنین، **PGD** در l_∞ از رویکردهای سابق قدرت بیشتری دارند. در بخش دیگری از این مقاله به بیان تعریف دفاع از شبکه‌های عصبی عمیق نسبت‌به نمونه‌های خصمانه و همچنین، به مرور و نقد برخی از مهمترین رویکردهای دفاع پرداختیم.

[1] Batch data

[DOI: 10.61186/jsdp.20.2.113]

رویکرد آموزش خصمانه را به‌عنوان رویکرد کلیدی و موفق جهت دفاع از حملات خصمانه معرفی کردیم. همچنین، بیان کردیم که تعریف صحیح عبارت خصمانه (عبارت موجود در تابع هزینه آموزش خصمانه) بسیار مهم است. در پایان، به نقد چند پژوهش در تحلیل فضای ویژگی فراگرفته‌شده در شبکه‌های عصبی عمیق پرداختیم و بر اساس پژوهش‌های موجود بیان کردیم که مصالحه میان تعمیم‌پذیری و صحت موثر بوده و الزاماً این دو مفهوم در تلاقی با یکدیگر نیستند. در بخش آخر نیز، به انجام آزمایش‌هایی جهت بررسی رویکردهای مختلف حمله خصمانه و بررسی قدرت هریک پرداختیم. بیان شد که در ناحیهٔ محصورشده $\boldsymbol{l}_\infty$ روش **PGD** و در ناحیه محصورشده $\boldsymbol{l}_2$ روش **DeepFool** قدرت بیشتری دارند. زمان اجرا از نکاتی بود که تحلیل، و مشخص شد که باتوجه به برتری روش‌های **PGD** و **DeepFool**، مدت زمان بیشتری نسبت‌به سایر هم‌ترازهای خود برای اجرا نیاز دارند و این میزان در **DeepFool** از همه روش‌های حمله بیشتر بود. همچنین، به مقایسهٔ برخی از رویکردهای پرکاربرد دفاع نسبت‌به نمونه‌های خصمانه نیز پرداخته‌شد و از بین روش‌های مبتنی‌بر نواحی حول داده $\boldsymbol{l}_\infty$، روش آموزش خصمانه مبتنی‌بر **PGD** با شاخص‌های مربوط از سایر روش‌ها بهتر بوده و در مقابل اغلب روش‌های حمله مقاوم عمل کرده‌است. گفتنی است که روش‌های مختلف حملهٔ خصمانه و همچنین، روش‌های مختلف دفاع نسبت‌به آن حملات که در این مقاله بررسی شد، برای استفاده عموم در نشانی **https://github.com/khalooei/ars** قابل استفاده است.

۷- مراجع

[1] I. Goodfellow, Y. Bengio, and A. Courville, Deep learning. MIT press, 2016.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[3] A. H. Marblestone, G. Wayne, and K. P. Kording, "Toward an integration of deep learning and neuroscience," Frontiers in computational neuroscience, vol. 10, p. 94, 2016.

[4] S. Ganguli, "Towards bridging the gap between neuroscience and artificial intelligence." [Online]. Available: https://cbmm.mit.edu/sites/default/files/documents/Ganguli_AAAI17_SoI.pdf. [Accessed: 01-Dec-2019].

[5] Y. LeCun and Y. Bengio, "The Handbook of Brain Theory and Neural Networks," M. A. Arbib, Ed. Cambridge, MA, USA: MIT Press, 1998, pp. 255–258.

[6] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "Adversarial Attacks and Defences: A Survey," 2018. [Online]. Available: http://arxiv.org/abs/1810.00069. [Accessed: 17-Aug-2019].

[7] A. D. Joseph, B. Nelson, B. I. P. Rubinstein, and J. D. Tygar, Adversarial machine learning. Cambridge University Press.

[8] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," in Proceedings of the International Conference on Learning Representations (ICLR), 2015.

[9] C. Szegedy et al., "Intriguing properties of neural networks," in Proceedings of the International Conference on Learning Representations (ICLR), 2014.

[10] A. Boloor, X. He, C. Gill, Y. Vorobeychik, and X. Zhang, "Simple Physical Adversarial Examples against End-to-End Autonomous Driving Models," in 2019 IEEE International Conference on Embedded Software and Systems (ICESS), 2019, pp. 1–7.

[11] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial Machine Learning at Scale," in Proceedings of the International Conference on Learning Representations (ICLR), 2017.

[12] N. Akhtar and A. Mian, "Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey," IEEE Access, vol. 6, pp. 14410–14430, 2018.

[13] S. Kariyappa and M. K. Qureshi, "Improving Adversarial Robustness of Ensembles with Diversity Training," 2019. [Online]. Available: http://arxiv.org/abs/1901.09981. [Accessed: 07-Oct-2019].

[14] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in Proceedings of the International Conference on Learning Representations (ICLR), 2017.

[15] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2574–2582.

[16] N. Carlini and D. Wagner, "Towards Evaluating the Robustness of Neural Networks," in Proceedings of the IEEE Symposium on Security and Privacy (SP), 2017, pp. 39–57.

[17] S. M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[18] J. Wu and R. Fu, "Universal, transferable and targeted adversarial attacks," 2019. [Online]. Available: http://arxiv.org/abs/1908.11332. [Accessed: 31-Dec-2019].

* نویسندهٔ عهده‌دار مکاتبات

* Corresponding author

• تاریخ ارسال مقاله: ۱۳۹۹/۱۰/۳۰ • تاریخ پذیرش: ۱۴۰۲/۴/۱۴ • تاریخ انتشار: ۱۴۰۲/۷/۳۰ • نوع مطالعه: کاربردی

[DOI: 10.61186/jsdp.20.2.113]

[19] Y. Dong et al., "Boosting Adversarial Attacks With Momentum," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[20] S. Qiu, Q. Liu, S. Zhou, and C. Wu, "Review of Artificial Intelligence Adversarial Attack and Defense Technologies," Applied Sciences, vol. 9, no. 5, p. 909, 2019.
[21] F. Assion et al., "The Attack Generator: A Systematic Approach Towards Constructing Adversarial Attacks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[22] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," Mathematical Programming, vol. 45, no. 1–3, pp. 503–528, 1989.
[23] T. Miyato, S.-I. Maeda, M. Koyama, and S. Ishii, "Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 8, pp. 1979–1993, 2019.
[24] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," in Proceedings of the International Conference on Learning Representations (ICLR), 2018.
[25] S. Sabour, Y. Cao, F. Faghri, and D. J. Fleet, "Adversarial Manipulation of Deep Representations," in Proceedings of the International Conferenceon Learning Representations (ICLR), 2016.
[26] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The Limitations of Deep Learning in Adversarial Settings," in Proceedings of the IEEE European Symposium on Security and Privacy (EuroS&P), 2016, pp. 372–387.
[27] J. Su, D. V. Vargas, and K. Sakurai, "One Pixel Attack for Fooling Deep Neural Networks," IEEE Transactions on Evolutionary Computation, vol. 23, no. 5, pp. 828–841, 2019.
[28] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks," in Proceedings of the IEEE Symposium on Security and Privacy (SP), 2016, pp. 582–597.
[29] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "ZOO: Zeroth Order Optimization Based Black-box Attacks to Deep Neural Networks without Training Substitute Models," in Proceedings of the ACM Workshop on Artificial Intelligence and Security (AISec), 2017, pp. 15–26.
[30] L. Rosasco, E. De Vito, A. Caponnetto, M. Piana, and A. Verri, "Are Loss Functions All the Same?," Neural Computation, vol. 16, no. 5, pp. 1063–1076, 2004.
[31] Z. Zhao, D. Dua, and S. Singh, "Generating Natural Adversarial Examples," Proceedings of the International Conference on Learning Representations (ICLR), 2018.
[32] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein Generative Adversarial Networks," in Proceedings of the International Conference on Machine Learning (ICML), 2017, vol. 70, pp. 214–223.
[33] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved Training of Wasserstein GANs," in Advances in Neural Information Processing Systems (NIPS), 2017, pp. 5767–5777.
[34] M. Arjovsky and L. Bottou, "Towards Principled Methods for Training Generative Adversarial Networks," Proceedings of the International Conference on Learning Representations (ICLR), 2017.
[35] T. Salimans et al., "Improved Techniques for Training GANs," in Advances in Neural Information Processing Systems (NIPS), 2016, pp. 2234–2242.
[36] M. Rosca, B. Lakshminarayanan, D. Warde-Farley, and S. Mohamed, "Variational Approaches for Auto-Encoding Generative Adversarial Networks," 2017. [Online]. Available: http://arxiv.org/abs/1706.04987.
[37] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial Examples: Attacks and Defenses for Deep Learning," IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 9, pp. 2805–2824, 2019.
[38] D. Stutz, M. Hein, and B. Schiele, "Disentangling Adversarial Robustness and Generalization," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[39] A. Mustafa, S. Khan, M. Hayat, R. Goecke, J. Shen, and L. Shao, "Adversarial Defense by Restricting the Hidden Space of Deep Neural Networks," in The IEEE International Conference on Computer Vision (ICCV), 2019.
[40] G. Tao, S. Ma, Y. Liu, and X. Zhang, "Attacks Meet Interpretability: Attribute-steered Detection of Adversarial Samples," in Advances in Neural Information Processing Systems 31, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 7717–7728.
[41] E. Wong, L. Rice, and J. Z. Kolter, "Fast is better than free: Revisiting adversarial training," 2019.
[42] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble Adversarial Training: Attacks and Defenses," Proceedings of the International Conference on Learning Representations (ICLR), 2018.
[43] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical Black-Box Attacks against Machine

[DOI: 10.61186/jsdp.20.2.113]

Learning," in Proceedings of the ACM on Asia Conference on Computer and Communications Security (ASIA CCS), 2017, pp. 506–519.

[44] H. Hosseini, Y. Chen, S. Kannan, B. Zhang, and R. Poovendran, "Blocking Transferability of Adversarial Examples in Black-Box Learning Systems," 2017. [Online]. Available: http://arxiv.org/abs/1703.04318.

[45] G. K. Dziugaite, Z. Ghahramani, and D. M. Roy, "A study of the effect of JPG compression on adversarial images," 2016. [Online]. Available: http://arxiv.org/abs/1608.00853. [Accessed: 01-Oct-2019].

[46] N. Das et al., "Keeping the Bad Guys Out: Protecting and Vaccinating Deep Learning with JPEG Compression," 2017. [Online]. Available: http://arxiv.org/abs/1705.02900. [Accessed: 01-Oct-2019].

[47] N. Akhtar, J. Liu, and A. Mian, "Defense Against Universal Adversarial Perturbations," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3389–3398.

[48] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille, "Adversarial Examples for Semantic Segmentation and Object Detection," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1378–1387.

[49] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial Examples Are Not Bugs, They Are Features," in Advances in Neural Information Processing Systems 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 125–136.

[50] D. Meng and H. Chen, "MagNet," in Proceedings of the ACM Conference on Computer and Communications Security (CCS), 2017, pp. 135–147.

[51] B. Biggio, B. Nelson, and P. Laskov, "Support Vector Machines Under Adversarial Label Noise," in Proceedings of the Asian Conference on Machine Learning, 2011, pp. 97–112.

[52] S.-I. Mirzadeh, M. Farajtabar, A. Li, and H. Ghasemzadeh, "Improved Knowledge Distillation via Teacher Assistant: Bridging the Gap Between Student and Teacher," in Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI), 2020.

[53] N. Carlini and D. Wagner, "Defensive Distillation is Not Robust to Adversarial Examples," eprint arXiv:1607.04311, 2016. [Online]. Available: http://arxiv.org/abs/1607.04311. [Accessed: 10-Oct-2019].

[54] W. Xu, D. Evans, and Y. Qi, "Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks," Network and Distributed Systems Security Symposium (NDSS) 2018, 2018.

[55] J. Gao, B. Wang, Z. Lin, W. Xu, and Y. Qi, "DeepCloak: Masking Deep Neural Network Models for Robustness Against Adversarial Samples," Proceedings of the International Conference on Learning Representations (ICLR), 2017.

[56] S. Gu and L. Rigazio, "Towards deep neural network architectures robust to adversarial examples," in Proceedings of the International Conference on Learning Representations (ICLR) Workshop, 2015.

[57] P. Samangouei, M. Kabkab, and R. Chellappa, "Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models," Proceedings of the International Conference on Learning Representations (ICLR), 2018.

[58] I. Goodfellow et al., "Generative Adversarial Nets," in Advances in Neural Information Processing Systems (NIPS), 2014, pp. 2672–2680.

[59] F. Liao, M. Liang, Y. Dong, T. Pang, X. Hu, and J. Zhu, "Defense Against Adversarial Attacks Using High-Level Representation Guided Denoiser," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2018.

[60] J. Uesato, B. O'Donoghue, P. Kohli, and A. Oord, "Adversarial Risk and the Dangers of Evaluating Against Weak Attacks," in Proceedings of Machine Learning Research, 2018, pp. 5025–5034.

[61] B. Sun, N.-H. Tsai, F. Liu, R. Yu, and H. Su, "Adversarial Defense by Stratified Convolutional Sparse Coding," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[62] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness May Be at Odds with Accuracy," Proceedings of the International Conference on Learning Representations (ICLR), 2018.

[63] D. Su, H. Zhang, H. Chen, J. Yi, P.-Y. Chen, and Y. Gao, "Is Robustness the Cost of Accuracy? – A Comprehensive Study on the Robustness of 18 Deep Image Classification Models," Springer, Cham, 2018, pp. 644–661.

[64] F. Tramèr, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "The Space of Transferable Adversarial Examples," 2017. [Online]. Available: http://arxiv.org/abs/1704.03453. [Accessed: 20-Oct-2019].

[65] X. Wang et al., "Adversarial Examples for Improving End-to-end Attention-based Small-footprint Keyword Spotting," in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2019, vol. 2019-May, pp. 6366–6370.

[66] L. Schönherr, K. Kohls, S. Zeiler, T. Holz, and D. Kolossa, "Adversarial Attacks Against Automatic Speech Recognition Systems via

[DOI: 10.61186/jsdp.20.2.113]

Psychoacoustic Hiding," in Network and Distributed System Security Symposium (NDSS), 2019.
[67] J. Ebrahimi, A. Rao, D. Lowd, and D. Dou, "Hotflip: White-box adversarial examples for text classification," in ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers), 2018, vol. 2, pp. 31–36.
[68] C. C. and C. B. Yann LeCun, "MNIST handwritten digit database." [Online]. Available: http://yann.lecun.com/exdb/mnist/. [Accessed: 24-Jun-2019].
[69] and G. H. Alex Krizhevsky, Vinod Nair, "CIFAR-10 and CIFAR-100 datasets," 2009. [Online]. Available: https://www.cs.toronto.edu/~kriz/cifar.html. [Accessed: 19-Oct-2019].
[70] S. Zagoruyko and N. Komodakis, "Wide Residual Networks," in Procedings of the British Machine Vision Conference 2016, 2016, vol. 2016-September, pp. 87.1-87.12.
[71] G. F. Silva, "CNN - Digit Recognizer (PyTorch) | Kaggle," Kaggle.com, 2018. [Online]. Available: https://www.kaggle.com/gustafsilva/cnn-digit-recognizer-pytorch. [Accessed: 14-Dec-2020].
[72] A. Paszke et al., "Automatic differentiation in PyTorch," 2017.
[73] R. (Roger) Fletcher, Practical methods of optimization. Wiley, 1987.

محمد خالوئی از سال ۱۳۹۶ به عنوان دانشجوی دکتری مهندسی کامپیوتر در گرایش هوش مصنوعی دانشگاه صنعتی امیرکبیر تحت نظارت دکتر محمدمهدی همایون‌پور و دکتر مریم امیرمزلقانی مشغول به تحصیل شد. زمینه‌های پژوهشی موردعلاقهٔ ایشان عبارت است از: یادگیری عمیق، یادگیری ماشین خصمانه، شبکه‌های مولد و بهینه‌سازی. نشانی رایانامهٔ ایشان عبارت است از:

khalooei@aut.ac.ir

محمد مهدی همایون‌پور در سال ۱۳۷۴ در رشته مهندسی برق از دانشگاه پاریس ۱۱ مدرک دکترای خود را اخذ و از آن تاریخ در گروه هوش مصنوعی دانشکده مهندسی کامپیوتر در دانشگاه صنعتی امیرکبیر مشغول به کار شد. زمینه‌های پژوهشی موردعلاقهٔ ایشان عبارت است از: پردازش گفتار، پردازش سیگنال، پردازش زبان طبیعی، یادگیری ماشین، یادگیری عمیق و همچنین، طراحی سخت‌افزار. نشانی رایانامهٔ ایشان عبارت است از:

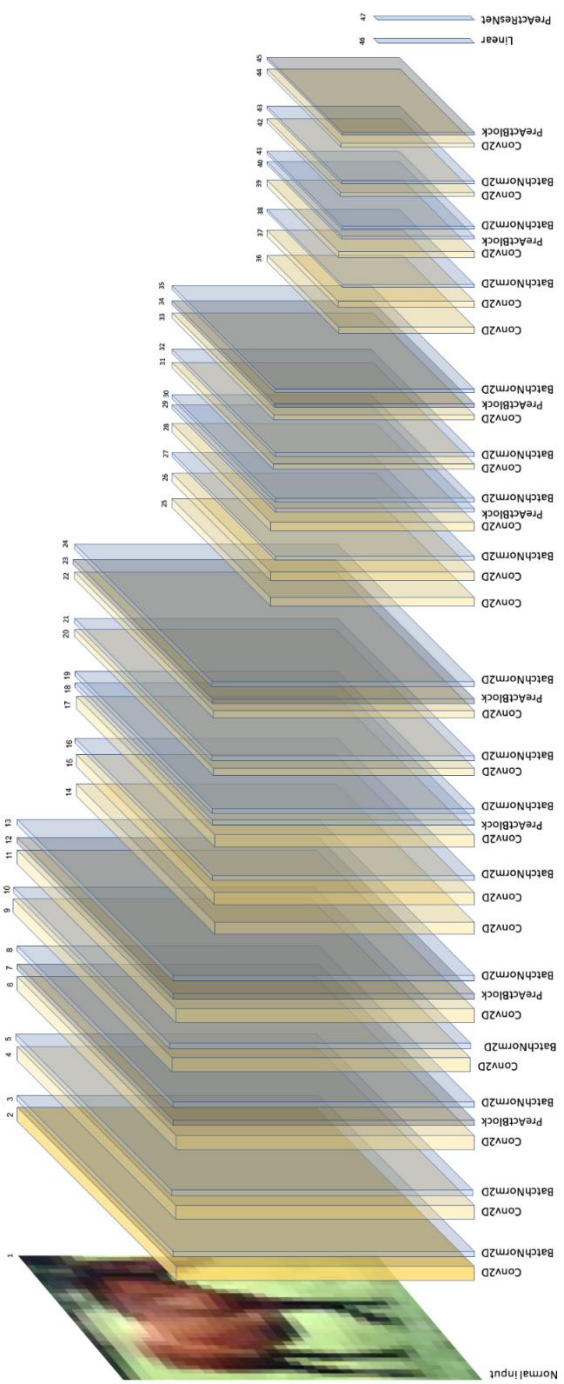
homayoun@aut.ac.ir

مریم امیرمزلقانی در سال ۱۳۸۸ در رشته مهندسی برق از دانشگاه صنعت ی امیرکبیر مدرک دکتری خود را دریافت کرد. وی هم‌اکنون عضو هیئت علمی گروه هوش مصنوعی در دانشکده مهندسی کامپیوتر دانشگاه صنعتی امیرکبیر است. زمینه‌های پژوهشی موردعلاقهٔ ایشان عبارت است از: پردازش سیگنال آماری، پردازش تصویر، پردازش سیگنال چندرزولوشنی و اخفای اطلاعات. نشانی رایانامهٔ ایشان عبارت است از:

mazlaghani@aut.ac.ir

[DOI: 10.61186/jsdp.20.2.113]

پیوست: نمایی از معماری شبکه **WRN-28**



[DOI: 10.61186/jsdp.20.2.113]