

# مروری بر آسیب‌پذیری شبکه‌های عصبی

## عمیق نسبت به نمونه‌های خصمانه و

## رویکردهای مقابله با آنها

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

### چکیده

امروزه شبکه‌های عصبی به عنوان بارزترین ابزار مطرح در هوش مصنوعی و یادگیری ماشین شناخته شده، و در حوزه‌های مالی و بانک‌داری، کسب و کار، تجارت، سلامت، پزشکی، بیمه، رباتیک، هواپیمایی، خودرو، نظامی و سایر حوزه‌ها استفاده می‌شوند. در سال‌های اخیر موارد بی‌شماری از آسیب‌پذیری شبکه‌های عصبی عمیق نسبت به حملاتی مطرح شده که به طور غالب، با افزودن اختلالات جمع‌شونده و غیرجمع‌شونده بر داده ورودی ایجاد می‌شوند. این اختلالات با وجود نامحسوس بودن در ورودی از دیدگاه عامل انسانی، خروجی شبکه آموزش‌دیده را تغییر می‌دهند. به اقداماتی که شبکه‌های عصبی عمیق را نسبت به حملات مقاوم می‌کنند، دفاع اطلاق می‌شود. برخی از روش‌های حمله، مبتنی بر ابزارهایی نظیر گرادینان شبکه نسبت به ورودی، در پی شناسایی اختلال هستند و برخی دیگر به تخمین آن ابزارها می‌پردازند و در تلاشند حتی بدون داشتن اطلاعاتی از آنها، به اطلاعاتشان دست پیدا کنند. رویکردهای دفاع نیز برخی روی تعریف تابع هزینه بهینه و همچنین، معماری شبکه مناسب، و برخی دیگر بر جلوگیری، یا اصلاح داده قبل از ورود به شبکه متمرکز می‌شوند. همچنین، برخی رویکردها به تحلیل میزان مقاوم بودن شبکه نسبت به این حملات و ارائه محدوده اطمینان متمرکز شده‌اند. در این مقاله سعی شده است جدیدترین پژوهش‌ها در زمینه آسیب‌پذیری شبکه‌های عصبی عمیق بررسی و نقد شوند، و با انجام آزمایش‌هایی نسبت به بررسی کارایی هریک، اقدام و آنها را با هم مقایسه شده‌اند. در آزمایش‌های انجام‌شده در بین حملات محصورشده به  $l_2$  و  $l_\infty$ ، به ترتیب روش PGD و روش DeepFool کارایی بالاتری دارند. زمان اجرا نیز از نکاتی است که تحلیل شده و نشان داده شده است که با توجه به برتری روش‌های PGD و DeepFool، این دو روش برای اجرا، مدت زمان بیشتری نسبت به سایر روش‌های هم‌ردیف خود نیاز دارند و این میزان در DeepFool از همه روش‌های حمله بیشتر است. همچنین، به مقایسه برخی از رویکردهای پرکاربرد دفاع نسبت به نمونه‌های خصمانه نیز پرداخته شد؛ که از بین روش‌های مبتنی بر نواحی محصورشده به  $l_\infty$  حول داده، روش آموزش خصمانه، مبتنی بر PGD با شاخص‌های معین، از سایر روش‌ها بهتر و در مقابل اغلب روش‌های حمله مقاوم بوده است. گفتنی است که روش‌های مختلف حمله خصمانه و همچنین، روش‌های مختلف دفاع نسبت به آن حملات که در این مقاله بررسی شده است، از طریق نشانی <https://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

<sup>1</sup>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

\* Corresponding author

\* نویسنده عهده‌دار مکاتبات

سال ۱۴۰۲ شماره ۲ پیاپی ۵۶

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



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  $l_\infty$  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\\_robustness\\_stack](https://github.com/khalooei/adversarial_robustness_stack).

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

## ۱- مقدمه

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

اندیشه استفاده از لایه‌های بیشتر در شبکه‌های عصبی و عمیق شدن آن‌ها از الهاماتی است که از علوم اعصاب<sup>۱</sup> برگرفته شده است [۳]–[۵]. شبکه‌های عصبی با افزایش لایه‌ها، فرآیند یادگیری و درک ویژگی‌های داده ورودی را بهتر انجام می‌دهند. بالا رفتن تعداد لایه‌ها و انواع آن‌ها در شبکه‌های عصبی عمیق موجب بیشتر شدن پیچیدگی و بزرگ شدن فضای پارامتری می‌شود. افزایش شاخص‌ها نیز به نوبه خود موجب طولانی و سخت شدن

جستجو در راستای رسیدن به پاسخ سراسری بهینه می‌شود [۱]، [۲] [۵]. طراحی ساختار و ارائه معماری شبکه‌های عصبی عمیق، تعداد لایه‌ها، نحوه چینش آن‌ها و سایر شاخصه‌های شبکه، امری پیچیده و بعضاً مبتنی بر تجربه است. همین امر موجب شده تا در کاربردهای حساس، استفاده از این الگوهای یادگیری قابل تأمل شود. درواقع، نباید تنها به دقت، صحت و سایر معیارهای ارزیابی مرسوم شبکه‌های عصبی عمیق توجه کرد؛ و نیاز است از نظر مقاوم بودن نسبت به حملات خصمانه نیز ارزیابی شوند [۶].

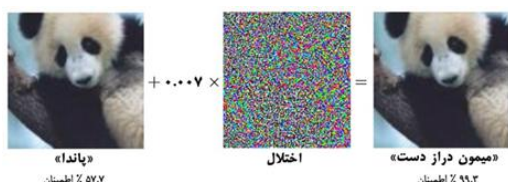
به بیان دیگر، کارایی بالای شبکه‌های عصبی عمیق و از طرفی استفاده فزاینده از آن‌ها، انگیزه‌ای را در افراد شورو و متخاصم ایجاد کرده است تا با دستکاری داده‌های ورودی به شبکه‌های عصبی عمیق الگو، عملکرد آن‌ها را مختل و در کار خود (مانند دسته‌بندی) دچار اشتباه کند، به این کار حمله خصمانه گفته می‌شود. حملات خصمانه و در نتیجه، تشخیص اشتباه توسط الگو ممکن است عواقب وخیمی را به دنبال داشته باشد. عدم تشخیص درست تابلوی «یست»

<sup>1</sup> Neuro science

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<sup>۳</sup>

به تغییرات بسیار اندک که در عین نامحسوس بودن از دیدگاه عامل انسانی، موجب اشتباه در تشخیص رده داده ورودی توسط الگوی یادگیری ماشینی می‌شود، اختلال<sup>۴</sup> و به اقداماتی که برای افزودن اختلال و به‌اشتباه‌انداختن الگو صورت می‌گیرد، حمله گفته می‌شود. حملات موجب فریب الگوی یادگیری ماشینی می‌شود و کلاس خروجی را به‌ازای داده ورودی تغییر می‌دهند. گفتنی است که با استفاده از حملات بی‌شمار به الگوهای یادگیری ماشینی، آسیب‌پذیری<sup>۵</sup> این الگوها به روش‌های مختلفی قابل‌ارزیابی است [۶]، [۸]، [۱۱] و [۱۸].



(شکل ۱- یک حمله نمونه با الگوی شبکه عصبی عمیق [۸])  
(Figure 1) A sample attack on the Deep Neural Network model [8]

برای مثال، در دسته‌بندی داده‌ها، اگر داده 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  $x$  با دقت ۵۷/۷ درصد «پاندا» تشخیص داده می‌شود. 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 پیدا کردن یک اختلال بهینه، نیازمند حل مسئله 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 رده یک داده خاص است و قابلیت عام‌بودن ندارد. اما می‌توان با روش‌هایی

توسط برنامه 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. همه اینها مثال‌هایی از اهمیت مقابله با حملات خصمانه و لزوم مقاوم‌سازی شبکه‌های عصبی عمیق نسبت به این حملات هستند.

یادگیری ماشینی تخصصی<sup>۱</sup> بر مقاوم‌سازی الگوهای یادگیری ماشینی نسبت به حملات تمرکز دارد. درصد پاسخ درست الگوهای یادگیری ماشینی با وجود حملات خصمانه، تعیین‌کننده میزان مقاومت آن‌ها نسبت به این گونه حملات است. متخصص در تلاش است تا با استفاده از داده مخرب<sup>۲</sup>، الگوهای یادگیری ماشینی، به‌ویژه شبکه‌های عصبی را فریب دهد.

مسائلی که مبتنی بر یادگیری ماشینی 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 تخصصی، 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 داده 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

<sup>3</sup> Attack

<sup>4</sup> Perturbation

<sup>5</sup> Vulnerability

<sup>1</sup> Adversarial Machine Learning

<sup>2</sup> Malicious



(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 در دسترس

### حمله جعبه سفید

در حمله جعبه سفید 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 حمله گریز یا حمله زمان تصمیم گیری است. اگر این اطلاعات تنها به جزئیات یادگیری نظیر داده آموزشی و موارد مشابه مربوط باشد، حمله از نوع حمله مسمومیت است. در حمله مسمومیت، با روش های گوناگونی نظیر تزریق داده آلوده به دادگان آموزشی تمیز، تغییر برچسب برخی از دادگان آموزشی و به طور کلی مختل کردن روال آموزش، می توان با تغییر در الگو بر خروجی نهایی تأثیر مخرب داشت. دسترسی متخاصم به اطلاعات کامل از الگو به معنی دسترسی او به اطلاعات دقیق شاخص های الگوی نظیر معماری، ویژگی های فراگرفته شده (وزن ها)، بهینه ساز و شاخص های الگوست [۶] و [۱۱].

### حمله جعبه سیاه

برخلاف حمله جعبه سفید، در این نوع حمله، متخاصم هیچ گونه اطلاع دقیقی از الگو یا فرآیندهای آموزشی نظیر آماده سازی داده و جزئیات آن در اختیار ندارد. در ساده ترین سطح، متخاصم می تواند به صورت محدود به الگو ورودی دهد و خروجی آن را دریافت کند. برای مثال، اگر ورودی  $x$  به عنوان یک درخواست به سامانه داده شود، الگو بر اساس ورودی درخواست شده، خروجی های متناظر را تولید می کند ( $y_{pred} = F(x)$ ). متخاصم در این صورت می تواند بر پایه  $y_{pred}$  دریافتی، تخمین هایی از  $F$  داشته باشد. در سطح بعد ممکن است متخاصم اطلاعاتی نظیر رده های فراگرفته شده در الگو، ویژگی ها و شاید اندکی از جزئیات یادگیری را در اختیار داشته باشد؛ اما اطلاعاتی از داده آموزشی نداشته باشد.

در خصوص محدودیت دسترسی به دادگان آموزشی ممکن است متخاصم هیچ گونه اطلاعاتی از الگو در اختیار نداشته باشد، ولی به تعدادی نمونه آموزشی که الگو با آن آموزش دیده، دسترسی داشته باشد. در نتیجه، زمانی که

اختلالی عام منظوره ای<sup>۱</sup> (همگانی) را تولید کرد که پس از اعمال روی مجموعه ای از داده ها موجب فریب الگو شود. صورت کلی مسئله حمله به صورت رابطه (۱) تبیین می شود:

$$\min_{\eta} \|\eta\|_p \quad s.t. \quad F(x + \eta) \neq F(x) \quad (1)$$

در رابطه (۱)، منظور از  $x$  داده ورودی،  $\eta$  همان اختلال و  $F$  دسته بند یا الگوی فراگرفته شده مسئله است. منظور از  $\|\cdot\|_p$  نیز نرم  $l_p$  است. حملات با استفاده از نظرات مختلف، اختلال ها را پیدا می کنند. در بخش های آینده انواع رویکردهای گوناگون برای یافتن اختلال بیان شده است [۷] و [۸].

## انواع حملات

دیدگاه های گوناگونی برای تقسیم بندی روش های مختلف حمله بیان شده است. در این پژوهش سعی شده در بهینه ترین حالت این تقسیم بندی ارائه شود. این تقسیم بندی از سه دیدگاه مرحله و زمان حمله، اطلاعات در دسترس متخاصم و اهداف کلی متخاصم انجام شده است.

### از دیدگاه مرحله وقوع حمله

یکی از مورد توجه ترین نکات در حمله، مرحله وقوع حمله است. حمله می تواند به صورت مستقیم روی الگو تأثیر داشته باشد. زمانی که حمله روی الگوی آموزش دیده انجام شود (مرحله آزمون)، به آن حمله گریز<sup>۲</sup> گفته می شود و یکی از مرسوم ترین حملات است. و زمانی که حمله با ایجاد تغییرات مخرب بر روی داده آموزشی (مرحله آموزش) صورت گیرد، به آن حمله مسمومیت<sup>۳</sup> گفته می شود. همان طور که در (شکل-۲) مشاهده می شود، حملات گریز، اغلب، در مرحله آزمون (روی الگوی  $F$ ) انجام می شوند. حملات مسمومیت برخلاف حملات گریز، در زمان آموزش و یا آماده سازی داده انجام می شوند.

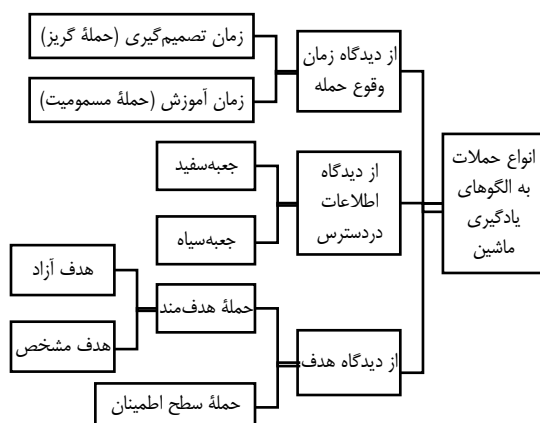
از دیدگاه اینکه متخاصم چه میزان داده و اطلاعات از سیستم مورد هدف در اختیار دارد، حملات به دو دسته جعبه سفید<sup>۴</sup> و جعبه سیاه<sup>۵</sup> تقسیم بندی می شوند. در ادامه به بیان تفصیلی این دو دسته حمله خواهیم پرداخت [۶] و [۱۱].

<sup>1</sup> Universal  
<sup>2</sup> Evasion  
<sup>3</sup> Poisoning  
<sup>4</sup> White-box  
<sup>5</sup> Black-box

سطح اطمینان پاسخ آن کاهش یابد و میزان خطای خروجی الگو نسبت به خروجی سابق تا حد امکان افزایش یابد. برای مثال، اگر سطح اطمینان دسته‌بند، پیشتر برای داده‌های ورودی از رده مد نظر ۹۸ درصد بوده باشد، 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 شده است.

#### تولید نمونه خصمانه<sup>۴</sup>

به‌منظور پیدا کردن و شناسایی نمونه خصمانه از الگوریتم‌ها و روش‌های متفاوتی استفاده می‌شود. در حالت کلی نیاز است متناسب با نوع حمله، اختلال را تولید، سپس، با داده 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 شده است. رابطه تولید نمونه خصمانه به‌صورت رابطه (۲) بیان می‌شود:

$$\hat{x} = x + \arg\min_{\eta} \{ \| \eta \| : F(x + \eta) \neq F(x) \} \quad (2)$$

در رابطه (۲)، تولید نمونه خصمانه از طریق حل مسئله بهینه‌سازی انجام می‌شود. یافتن کمینه یک تابع هزینه دارای رویکردهای متنوعی است. در توابع محدب به‌راحتی می‌توان به کمینه محلی رسید و از آنجا که کمینه‌های محلی در توابع محدب، کمینه‌های سراسری نیز هستند، به‌نسبت و به‌راحتی می‌توان جواب را یافت. اما نکته اصلی اینجاست که تابع هزینه شبکه‌های عصبی و به‌ویژه، شبکه‌های عصبی عمیق، به‌طور عموم، از نوع

<sup>۴</sup> Adversarial example

متخاصم اطلاعات تقریبی یا دقیق از دادگان آموزشی داشته باشد، حمله می‌تواند به نوعی از حمله جعبه سفید تبدیل شود؛ زیرا در این صورت، متخاصم قادر خواهد بود با استفاده از دادگان آموزشی در دسترس، الگویی مشابه الگوی اصلی آموزش دهد [۶] و [۱۱].

اغلب بیان می‌شود که حملات جعبه سیاه به اصل انتقال‌پذیری متکی هستند [۱۳]. این بدان معنی است که می‌توان با استفاده از نمونه خصمانه یک الگو، الگوی دیگری را هدف قرار داد. درواقع، باتوجه به محدودیت دسترسی به الگو در حملات جعبه سیاه، اغلب تلاش می‌شود تا الگویی جایگزین<sup>۱</sup> به گونه‌ای آموزش ببیند که دقت و میزان پاسخ‌گویی آن به الگوی اصلی نزدیک باشد؛ سپس، با حمله جعبه سفید 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. باتوجه به اهداف حمله، دسته‌بندی گوناگونی از حملات، قابل ارائه است. اما به‌طور کلی می‌توان اهداف مختلف یک متخاصم را در دو دسته حملات هدف‌مند<sup>۲</sup> و حملات کاهش سطح اطمینان<sup>۳</sup> خلاصه کرد.

#### حملات هدف‌مند

در حملات هدف‌مند، هدف فریب‌دهنده این است که پاسخ الگو را از یک رده به رده دیگر تغییر دهد. این حمله خود، می‌تواند به دو دسته آزاد و مشخص تقسیم‌بندی شود. در حمله هدف آزاد، مقصود این است که رده داده ورودی به غیر از رده اصلی خودش تغییر یابد. در هدف مشخص، مقصود این است که رده داده ورودی به رده خاصی تغییر یابد. همچنین، حالت خاص‌تری وجود دارد و آن این است که رده مبدأ مشخصی به رده 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 سطح اطمینان

در حملات سطح اطمینان برخلاف حملات هدف‌مند، هدف آن است که با افزایش خطای پیش‌بینی الگو،

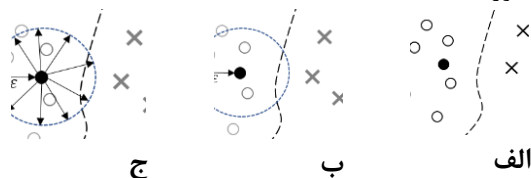
<sup>۱</sup> Surrogate

<sup>۲</sup> Targeted attack

<sup>۳</sup> Confidence reduction attack

$$\hat{x} = x + \varepsilon \cdot \text{sign}(\nabla_x J(\theta, x, y_{\text{true}})) \quad (5)$$

در عبارت بالا، منظور از  $\text{sign}$  تابع علامت،  $J$  تابع هزینه تعریف شده برای الگوی یادگیری،  $y_{\text{true}}$  برچسب متناسب با داده ورودی  $x$  و  $\varepsilon$  میزان تأثیر اختلال روی داده ورودی  $x$  در روش FGSM است. تعبیر هندسی مسئله بالا در شکل (۴) به صورت گام به گام از چپ به راست آورده شده است:



(شکل - ۴) تعبیر رویکرد FGSM با اعمال تغییرات به میزان  $\varepsilon$  به صورت مرحله به مرحله تا رسیدن به نمونه خصمانه مطلوب

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

طبق شکل (۴)، مبتنی بر نمونه‌ها، در قسمت الف یک نمونه انتخاب می‌شود، سپس، مبتنی بر نمونه انتخاب شده، این روش قصد دارد در محدوده  $\varepsilon$  به دنبال تولید اختلالی باشد که اگر به داده اضافه شود، (شکل - ب) موجب شود، نمونه تولیدی از مرز بین رده‌ای عبور کند (شکل (۴) - پ)؛ و آن نمونه در محدوده  $\varepsilon$  به عنوان نمونه خصمانه اعلام شود.

در نگاه آنان دلیل اصلی وجود نمونه خصمانه، فرض امکان جداسازی کلاسه‌ها به صورت خطی در فضای با ابعاد بالا بوده است. تعریف مسئله آن‌ها با نرم بی‌نهایت (استفاده از تابع علامت) بیان شده است. میاتو و همکاران تعریف مسئله را با استفاده از طبیعی سازی نرم  $l_2$  به فرم رابطه (۶) بیان کردند [۲۳]:

$$\eta = \varepsilon \frac{\nabla J(\theta, x, y)}{\|\nabla J(\theta, x, y)\|_2} \quad (6)$$

در رابطه (۶) سعی شده است تا گرادیان با نسبت نرم  $l_2$  طبیعی شود. حال این نرم اگر نرم بی‌نهایت باشد، به همان تعریف FGSM نسخه اولیه  $\eta = \varepsilon \cdot \text{sign}(\nabla J(\theta, x, y))$  تبدیل می‌شود.

**روش علامت گرادیان سریع هدفمند<sup>۷</sup>:** کوراکین<sup>۸</sup> و همکاران [۱۱]، قابلیت هدفمند شدن نمونه خصمانه را با کمی تغییر در مسئله بهینه سازی FGSM ارائه کردند. منظور از هدفمند بودن آن است که فریب صورت گرفته به گونه‌ای باشد تا برچسب رده به برچسب

غیرخطی و غیرمحدب هستند. این موضوع موجب می‌شود پیدا کردن کمینه به راحتی صورت نپذیرد. همچنین، بدیهی است که کمینه‌های محلی در توابع غیرمحدب، به الزام کمینه‌های سراسری نیستند و همین امر موجب سخت شدن و عدم تعریف فرم بسته برای این نوع از مسئله‌ها می‌شود [۷] و [۸].

### روش‌های مختلف تولید نمونه خصمانه

حل مسئله بهینه سازی رابطه (۲)، به روش‌های متعددی قابل انجام است. در ادامه سعی شده است تا به صورت مختصر مروری بر برخی پژوهش‌های مرتبط و کلیدی صورت پذیرد.

**روش L-BFGS<sup>۱</sup>:** ستردی<sup>۲</sup> و همکاران در پژوهش [۹] پیدا کردن اختلال بهینه برای تغییر رده ورودی به رده‌ای غیر از رده اصلی را با رابطه (۳) تعریف کردند:

$$\text{minimum } \|\eta\|_2 \quad (3)$$

$$\text{s.t. } F(x + \eta) \neq F(x) \\ x + \eta \in [0, 1]^m$$

رابطه بالا را می‌توان به صورت رابطه (۴) نیز بازنویسی کرد:

$$\text{minimum } c\|\eta\|_2 + \text{loss}_{F,t}(x + \eta) \quad (4) \\ \text{s.t. } x + \eta \in [0, 1]^m$$

در رابطه (۴) منظور از ثابت  $c$  ضریب ترکیب خطی است. همچنین، منظور از  $\text{loss}_{F,t}$  تابع هزینه دسته‌بند  $F$  مدنظر به صورت هدفمند با هدف مشخص  $t$  است. آن‌ها در پژوهش خود به منظور حل این مسئله از الگوریتم بهینه سازی L-BFGS<sup>۳</sup> [۲۲] استفاده کردند. آن‌ها همچنین، به حالت‌های خاصی از آموزش خصمانه<sup>۴</sup>، همچون افزودن نمونه‌های خصمانه به مجموعه داده و سپس، آموزش الگو نیز اشاره کردند. گفتنی است که روش پیدا کردن نمونه خصمانه آن‌ها دارای هزینه محاسباتی زیادی است.

**روش علامت گرادیان سریع (FGSM)<sup>۵</sup>:** گودفلو<sup>۶</sup> و همکاران در پژوهش [۸]، راه حل بهتری برای حل مسئله پیدا کردن نمونه خصمانه معرفی کردند. این رویکرد مبتنی بر اعمال تابع علامت روی گرادیان تابع هزینه نسبت به ورودی  $x$  است. نمونه خصمانه  $\hat{x}$  پژوهش آن‌ها توسط رابطه (۵) تولید می‌شود:

<sup>1</sup> Limited-Broyden Fletcher Goldfarb Shannon

<sup>2</sup> Szegedy

<sup>3</sup> BFGS نسخه بهینه شده و محدود شده حافظه‌ای روش بهینه سازی

[۷۳] می‌باشد.

<sup>4</sup> Adversarial training

<sup>5</sup> Fast Gradient Sign Method (FGSM)

<sup>6</sup> Goodfellow

<sup>7</sup> Targeted Fast Gradient Sign Method

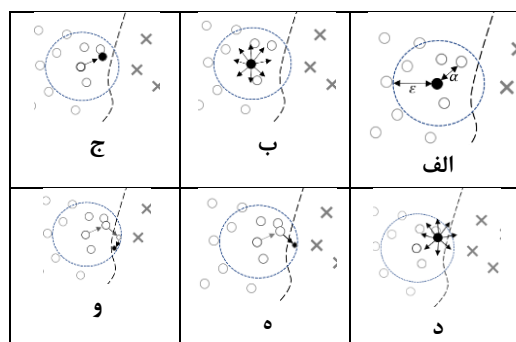
<sup>8</sup> Kurakin

تولیدشده، توسط عملیات  $\{Clip_{x,e}\}$ ، به محدوده  $[x_{i,j} - \epsilon, x_{i,j} + \epsilon]$  بازگردانی می‌شود. در (شکل - ۴) مراحل گام به گام الگوریتم iFGSM آورده شده‌است. قسمت الف (شکل - ۴)، نشان‌دهنده نسبت اندازه متغیرهای  $\alpha$  و  $\epsilon$  در فضای ویژگی مورد بحث است. در بخش‌های ب تا ج، روال پیشرفت الگوریتم تکرارشونده نشان داده شده‌است. این روش به نسبت رویکرد FGSM، نمونه خصمانه بهتری تولید می‌کند. کوراکین و همکاران در پژوهش خود در خصوص مقاوم‌سازی نیز بحث و به روش آموزش خصمانه پژوهش سزدی و همکاران [۹] نیز استناد کردند. نوع دیگری از رویکرد تکرارشونده iFGSM، روش تکرارشونده کم‌احتمال‌ترین رده<sup>۶</sup> است. کوراکین و همکاران [۱۱] آن را با نام ILLCM ارائه کردند که در رابطه Error! Reference source not found. آورده شده‌است. در این رابطه، رده یا کمترین احتمال با برچسب  $y_{LL}$  مشخص شده‌است.

(۹)

$$\hat{x}_0 = x$$

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



(شکل - ۴) گام‌های طرح‌واره الگوریتم iFGSM به صورت

مرحله به مرحله. شکل (الف) نشان‌دهنده نسبت اندازه

متغیرهای  $\alpha$  و  $\epsilon$  در فضای ویژگی و شکل‌های «ه» تا «و» روال پیشرفت الگوریتم تکرارشونده را تا رسیدن به نمونه خصمانه مطلوب نشان می‌دهد.

(Figure 5) Schematic steps of the iFGSM algorithm. Figure A in the figure shows the size ratios of the variables  $\alpha$  and  $\epsilon$  space and the other steps (figures) show the iterative progress until reaching the desired adversarial example

مَدری<sup>۸</sup> و همکاران در [۲۴] مسئله پیداکردن اختلال<sup>۹</sup> را با استفاده از روش کمینه گرادیان تصویرشده<sup>۱۰</sup>

<sup>۶</sup> Least-likely (LL) class method

<sup>۷</sup> Iterative Least-Likely Class method

<sup>۸</sup> Madry

<sup>۹</sup> مسئله پیداکردن اختلال بهینه، یک مسئله بهینه‌سازی دارای قید است.

<sup>۱۰</sup> Projected Gradient Descent

هدف موردنظر تغییر یابد. رابطه ریاضی بهینه‌سازی به صورت رابطه (۷) تعریف می‌شود:

$$\hat{x} = x - \epsilon \cdot sign(\nabla_x J(\theta, x, y_{target})) \quad (۷)$$

رابطه (۷) همانند رابطه (۵) است، با این تفاوت که در زمان محاسبه گرادیان، تغییرات برچسب هدف مسئله  $(y_{target})$  نسبت به تغییرات ورودی  $x$  محاسبه می‌شود. همچنین، پس از اعمال تابع علامت، این بار از  $x$  کم می‌شود. درواقع، برای این است تا بیشینه احتمال خروجی الگو برای رده  $y_{target}$  نسبت به ورودی نمونه خصمانه افزایش یابد.

### روش‌های تکرارشونده مبتنی بر گرادیان:

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

$$\hat{x}_0 = x \quad (۸)$$

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

در رابطه (۸) برای این که دست‌یابی به پاسخ نهایی در چندین گام صورت می‌پذیرد، اولین گام با مقداردهی نمونه اصلی  $(\hat{x}_0 = x)$  شروع می‌شود. در گام‌های بعدی بر مبنای آخرین نمونه مرحله قبل  $(\hat{x}_n)$ ، نمونه جدید  $(\hat{x}_{n+1})$  تولید می‌شود که در هر گام، FGSM را مبتنی بر آخرین نمونه مرحله قبل اجرا نماید. سپس، در صورتی که محدوده نمونه خصمانه تولیدی از محدوده تعریف‌شده<sup>۴</sup> عبور کرده باشد، مقادیر توسط عملیات  $Clip_{x,e}$  به محدوده موردانتظار برمی‌گردند<sup>۵</sup>. این مراحل با تعداد گام مشخص یا تا زمان رسیدن به نمونه خصمانه بهتر ادامه می‌یابد.

در رابطه (۸) منظور از  $\alpha$  اندازه گام حرکت و  $\{Clip_{x,e}\}$  عملیات محدودکردن ورودی به محدوده مشخص شده است. گفتنی است که هریک از اجزای نمونه

<sup>۱</sup> Iterative Fast Gradient Sign Method

<sup>۲</sup> Iterative FGSM (iFGSM)

<sup>۳</sup> Basic Iterative Method (BIM)

<sup>۴</sup> محدوده طبیعی مقادیر ماتریس عددی هر نمونه، به عنوان مثال،

محدوده طبیعی صفر و یک یا محدوده غیرطبیعی داده ورودی همانند محدوده [۰، ۲۵۵] در تصویر است.

<sup>۵</sup> همه عناصر ماتریس داده و یا همه پیکسل‌ها در داده‌های تصویری به محدوده مقادیر مشخص شده محدود و منتقل می‌شوند.

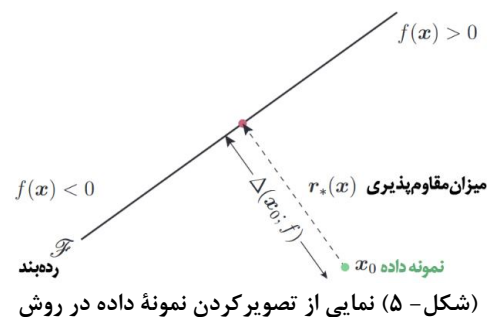
به جای گرادین کمینه حل کردند. در روش پیشنهادی آن‌ها، مراحل پیشروی همانند گرادین سریع تکرارشونده انجام می‌شود و تفاوت زیادی در طول روند پیشروی الگوریتم کلی وجود ندارد.

**دستکاری تخصصی بازنمایی عمیق<sup>۱</sup>:** صبور و همکاران در [۲۵]، روش حمله هدفمندی را ارائه کردند که در آن به جای دخیل کردن فاصله خروجی حاصل از الگوی (برچسب)، از فاصله خروجی حاصل از لایه‌های میانی استفاده کردند. درواقع، هدف آن‌ها با ارائه این نظر این بود که بیان کنند نباید اختلافی بین ویژگی‌های استخراج‌شده لایه میانی داده تمیز و داده خصمانه ایجاد شود. رابطه مسئله بهینه‌سازی آن‌ها در رابطه (۹) آورده شده است:

$$\min_{\hat{x}} \|\phi_k(x) - \phi_k(\hat{x})\| \quad (9)$$

$$s.t. \quad \|x - \hat{x}\|_{\infty} < \delta$$

در رابطه (۹) منظور از  $\phi_k$ ، نگاشتی از ورودی به خروجی لایه  $k$  ام شبکه است. همچنین، همانند بیان‌های مختلف دیگر، از مسئله پیدا کردن کمترین اختلال، نرم بی‌نهایت را برای محدود کردن 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 تصویر کردن نمونه 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  $l_2$ ، پیدا کردن اختلال را بهتر از روش‌های پیشین نظیر روش LBFGS انجام می‌دهد (در روش LBFGS تنها به استفاده از نرم  $l_2$  در محدود کردن تولید اختلال تأکید دارد). دیدگاه روش

آن‌ها با فرض تقریب خطی شبکه عصبی پیش می‌رود که در (شکل - ۵) با  $f$  نشان داده شده است. در نگاه آن‌ها تفکیک یک رده از رده دیگر توسط ابرصفحه انجام می‌شود. درواقع، آن‌ها با تصویر کردن<sup>۲</sup> نمونه به دسته‌بند سعی می‌کنند تا میزان مقاومت‌پذیری را به‌دست آورند و از آن برای تولید اختلال استفاده کنند. (شکل - ۵) نمایی از تصویر کردن نمونه داده بر دسته‌بند و گام‌های نخستین روش DeepFool را نشان می‌دهد.

به‌منظور تصویر کردن یک نقطه روی دسته‌بند، از رابطه (۱۰) استفاده می‌شود:

$$\eta \sim r_* := \operatorname{argmin} \|r\|_2 \quad (10)$$

$$s.t. \quad \operatorname{sign}(f(x_0 + r)) \neq \operatorname{sign}(f(x_0))$$

$$= -\frac{f(x_0)}{\|\nabla f(x_0)\|_2^2} \nabla f(x_0)$$

در رابطه (۱۰) منظور از  $r$  میزان مقاوم بودن الگو و به عبارت دیگر،  $r$  تعداد قدم‌های موردنیاز جهت رسیدن به مرز دسته‌بند است. با فرضیه‌های بیان‌شده از مسئله، آن‌ها راه حل 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 نمونه تمیز شروع می‌کند. در گام‌های بعدی متناسب با تقریب خطی دسته‌بند، سعی می‌کند روی آن تصویر<sup>۳</sup> شود. همین روال تا رسیدن به نمونه خصمانه نیز تکرار می‌شود. با جمع همه  $r_i$  ها، می‌توان به اندازه اختلال به دسته‌بند نزدیک شد.

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

$$\eta = (1 + \varepsilon) \times \hat{r} \quad (11)$$

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 دورده‌ای تصویرسازی و ادامه روال انجام می‌پذیرد.

**روش نقشه برجستگی مبتنی بر ژاکوبین<sup>۴</sup>:** اغلب روش‌های تولید نمونه خصمانه برای تولید اختلال‌های نامحسوس، با نرم‌های  $l_2$  یا  $l_{\infty}$  محدود می‌شوند. پپرنو<sup>۵</sup> و همکاران در پژوهش [۲۶]، اختلال را به نرم  $l_0$  محدود کردند. آن‌ها به جای تغییر کل مقادیر (پیکسل‌ها)، به دنبال

<sup>2</sup> Project

<sup>3</sup> Projection

<sup>4</sup> Jacobean Based Saliency Map (JBSM)

<sup>5</sup> Pepernot

<sup>1</sup> Adversarial manipulation of deep representations (Rep. Adversary)

**روش کارلینی و وگنر C&W<sup>4</sup>:** پپرنو و همکاران در پژوهش [۲۸]، خالص‌سازی دفاعی<sup>۵</sup> را به‌عنوان راه حلی جهت مقاوم‌شدن در مقابل نمونه‌های خصمانه ارائه کردند. کارلینی و وگنر در [۱۶] به‌صورت ویژه به ارائه سه رویکرد حمله در پاسخ به پژوهش پپرنو و همکاران [۲۸] پرداختند. رویکرد حمله پپرنو و همکاران، اختلال را با محدودیت نرم  $l_0$ ،  $l_1$  و  $l_\infty$  به‌دست‌می‌آورد. آن‌ها نشان دادند که روش خالص‌سازی دفاعی برای دفاع در مقابل نمونه‌های خصمانه مفید نیست. آن‌ها در نخست طرح کلی مسئله پیدا کردن نمونه خصمانه را در رابطه (۱۲) بیان کرده، سپس، برای هر بخش از آن، معادل‌هایی را پیشنهاد دادند که در نهایت به پاسخ بهینه رابطه ختم شود.

در رابطه (۱۲)، منظور از  $Dist$  معیار فاصله است<sup>۶</sup>:

$$\min Dist(x, x + \eta) \quad (12)$$

$$s. t. F(x + \eta) = y_{target}$$

$$x + \eta \in [0, 1]^n$$

آن‌ها بیان کردند با توجه به این که شرط  $F(x + \eta) = y_{target}$

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،  $f$  نامیدند و اقدام به جایگزین کردن آن با شرط مسئله کردند. به عبارت دیگر، شرط یادشده در صورتی صادق است که بتوان یک تابع مانند  $f$  پیدا کرد که در رابطه (۱۴) صدق کند.

$$F(x + \eta) = y_{target} \Leftrightarrow f(x + \eta, t) \leq 0 \quad (14)$$

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

$$\text{minimize} \left\| \frac{1}{2} (\tanh(w) + 1) - x \right\|_2^2 + c \quad (15)$$

$$\cdot \text{loss}_{f, y_{target}} \left( \frac{1}{2} (\tanh(w) + 1) \right)$$

$$s. t. f(\hat{x}) = \max(\max\{[Z(\hat{x})]_i; i \neq t\} - [Z(\hat{x})]_t, -\kappa)$$

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

کمترین تغییر مقدار (تغییر پیکسل) در داده (تصویر) ورودی بودند. در روش آن‌ها به محاسبه ژاکوبین  $F$  پرداخته شده‌است. آن‌ها با محاسبه گرادیان خروجی نسبت به ورودی در تلاش هستند تا اجزای تأثیرگذار و نقاط برجسته و بارز در داده 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 داده 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 محاسبه امتیاز تأثیرگذاری برجستگی ابعاد داده زمان‌بر است، اما مزیت آن این است که تغییرات بسیار اندکی روی داده ورودی اعمال می‌شود.

در روش آن‌ها، نخست نقشه برجستگی هر یک از عناصر<sup>۱</sup> ورودی توسط رابطه (۱۳) محاسبه می‌شود.

$$S(x, t)[i] = \begin{cases} 0, & \frac{\partial F_t}{\partial x_i}(x) < 0 \text{ or } \sum_{j \neq t} \frac{\partial F_j}{\partial x_i}(x) > 0 \\ \left| \frac{\partial F_t}{\partial x_i}(x) \right| \left| \sum_{j \neq t} \frac{\partial F_j}{\partial x_i}(x) \right|, & \text{otherwise} \end{cases} \quad (13)$$

در رابطه (۱۳)، منظور از  $S(x, t)[i]$  مقدار برجستگی عنصر  $i$ ام از داده  $x$  برای برجسب رده هدف  $t$  است. منظور از  $F_t$  خروجی تابع هزینه مبتنی بر رده 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<sup>۲</sup> و همکاران در [۲۷] از روش بهینه‌سازی تکامل تفاضلی<sup>۳</sup> برای به‌دست‌آوردن پاسخ مسئله پیدا کردن اختلال بهینه استفاده کردند. آن‌ها در تلاش بودند تا در الگوریتم‌شان با تغییر یک پیکسل، برجسب رده ورودی را تغییر دهند. روند پیشروی آن‌ها نیز مبتنی بر اصول پردازش‌های تکاملی، تولید جمعیت و روال مشابه در فضای بهینه‌سازی تکاملی است. نظر سو و همکاران در دسته‌بندی فعالیت‌های اختلال محدود به نرم  $l_0$  قرار می‌گیرد. نکته 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 برجسب‌های پیش‌بینی شده). بنابراین این روش جزء حملات جعبه‌سیاه نیز قرار می‌گیرد.

<sup>۱</sup> در داده‌های تصویری منظور از عنصر، پیکسل‌های تصویر است.

<sup>۲</sup> Su

<sup>۳</sup> Differential Evolution

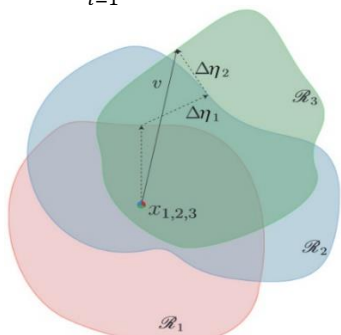
<sup>۴</sup> Carlini and Wanger

<sup>۵</sup> Defensive distillation

<sup>۶</sup> همانند بیشتر روش‌هایی که تاکنون تحت عناوین نرم‌های مختلف  $l_0$ ،  $l_2$  و  $l_\infty$  به‌عنوان معیار فاصله بیان شد.

صفر و شعاع  $\xi$  تصویر می کنند. به عبارت دیگر، این بیان در قالب رابطه (۱۸) بیان می شود:

$$Err(X_\eta) = \frac{1}{m} \sum_{i=1}^m 1_{F(x_i+\eta) \neq F(x_i)} \geq 1 - \delta \quad (18)$$

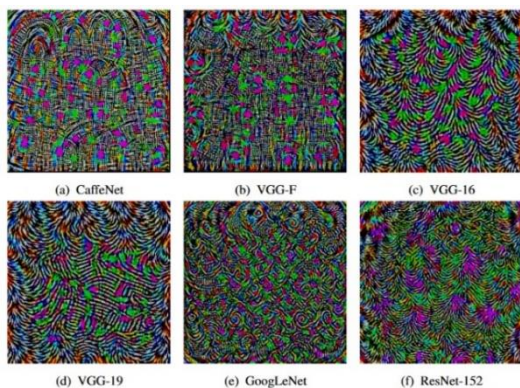


(شکل - ۶) طرح بازنمایی الگوریتم اختلال خصمانه فراگیر [۱۷]

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

در (شکل - ۶)، خلاصه ای از مراحل الگوریتم اختلال خصمانه فراگیر به تصویر کشیده شده است. در این شکل نمونه داده های  $x_1$ ،  $x_2$  و  $x_3$  در نمای دوبعدی به طور تقریب، روی هم افتاده اند. محدوده های هر دسته بند با استفاده از هاله های رنگی  $\mathcal{R}_1$ ،  $\mathcal{R}_2$  و  $\mathcal{R}_3$  نشان داده شده است.

الگوریتم روش آن ها، در هر تکرار، اختلال کمینه را تجمع کرده و سعی می کند داده مختل شده را به خارج از محدوده  $\mathcal{R}_i$  هدایت کند. همین روال به صورت تکرارشونده تا خارج شدن از نواحی مختلف و به میزانی که کاربر درخواست کرده، ادامه می یابد.



(شکل - ۷) - نمونه ای از اختلال های فراگیر محاسبه شده

برای معماری های مختلف شبکه های عصبی [۱۷] (برای نمایش بهتر مقادیر پیکسل ها تغییر مقیاس شده اند).

(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) نمونه ای از اختلال های فراگیر محاسبه شده برای معماری های مختلف شبکه های عصبی ترسیم شده است. در بسیاری از این اختلال ها، الگوهای قابل توجهی مشاهده می شود. از طرفی در شبکه های مختلف الگوهای مشابه نیز مشاهده می شود.

خالص سازی شده در قبال حمله  $l_2$  مقاوم نبوده و قابل نفوذ است. منظور  $Z$  لاجیت (خروجی لایه آخر شبکه قبل از اعمال softmax) و  $w$  متغیر جایگزین  $\eta$  در  $Dist(x, x + \eta)$  و رویکرد آن ها جزاً روش های تکرارشونده است.

**روش اختلال خصمانه فراگیر<sup>۱</sup>:** اغلب روش های حمله بیان شده، اختلال را متناظر با هر نمونه داده به دست می آورند. موسوی دزفولی و همکاران در [۱۷]، تلاش کردند تا رویکردی جهت تولید اختلال خصمانه فراگیر ارائه کنند. با استفاده از اختلال خصمانه فراگیر، به جای این که اختلال مدنظر به نمونه داده وابسته باشد، به شبکه (الگو) وابسته است. درواقع، زمانی که اختلال فراگیر روی تعدادی از نمونه داده های ورودی اعمال شود، موجب می شود تا حمله خصمانه با موفقیت انجام شود و رده نمونه های متناظر تغییر یابد. گفتنی است که این اختلال نیز باید شرط نامحسوس بودن را از دیدگاه انسانی داشته باشد، ولی آن ها نیز اشاره کردند که اختلال یادشده شبه نامحسوس<sup>۲</sup> است. آن ها همچنین، اشاره کردند الزامی بر فراگیر بودن اختلال برای همه تصاویر نبوده و تنها اختلالی مدنظر است که بیشینه تصاویر را مختل کند و سامانه را فریب دهد. آن ها شروطی جهت پیدا کردن اختلال کمینه معرفی کردند که به صورت روابط (۱۶) بیان شده است:

$$\|\eta\|_p \leq \xi \quad (16)$$

$$\mathbb{P}_{x \sim X}(F(x + \eta) \neq F(x)) \geq 1 - \delta$$

در روابط (۱۶)، منظور از  $\mathbb{P}$  احتمال و منظور از کل عبارت، احتمال وجود اختلال روی تعدادی از نمونه ها نظیر  $x$  است. همچنین، منظور از  $\xi$  شاخص کنترل کننده میزان اختلال و  $\delta$  تعیین کننده نرخ فریب نمونه های نمونه برداری شده از توزیع  $X$  است. آن ها بیان کردند که اگر اختلال فراگیر روی مجموعه داده اندک محاسبه شود، نیاز است تا نرخ فریب، بزرگ انتخاب شود. گفتنی است که روش آن ها از روش های تکرارشونده بوده، و در هر تکرار کوچک ترین اختلال محاسبه شده و با مقدار جاری آن در آن مرحله، جمع بسته می شود. مسئله بهینه سازی آن ها به صورت رابطه (۱۷) بیان شده است:

$$\Delta \eta \leftarrow \arg \min_r \|r\|_2 \quad (17)$$

$$s. t. F(x_i + \eta + r) \neq F(x_i)$$

آن ها به منظور محدود کردن شرط رابطه (۱۷)، اختلال به دست آمده را در هر تکرار روی کره  $l_p$  با مرکزیت

<sup>1</sup> Universal Adversarial Perturbation

<sup>2</sup> Quasi imperceptible

تولید نمونه خصمانه طبیعی<sup>۴</sup>: ژائو<sup>۵</sup> و همکاران در [۳۱] تلاش کردند تا با استفاده از چارچوب شبکه تقابلی مولد<sup>۶</sup> نمونه خصمانه تصویری و متنی تولید کنند. نمونه‌های تولیدشده توسط پژوهش آن‌ها برای انسان نیز طبیعی به نظر می‌رسد. آن‌ها با استفاده از آموزش شبکه تقابلی مولد با تابع هزینه واسرشتین<sup>۷</sup> به نام شبکه تقابلی مولد واسرشتین (WGAN)<sup>۸</sup> [۵۴] و [۵۵] این فعالیت را انجام دادند. همچنین، آن‌ها یک شبکه معکوس<sup>۹</sup>  $J$  را در کنار شبکه تقابلی مولد اصلی قرار دادند تا نمونه‌های خصمانه را به فضای نهان هدایت کند. طبق بیان آن‌ها خطای بازسازی نمونه  $x$  همراه با اختلاف بین توزیع گوسی  $z$  و  $J_r(G_\theta(z))$  توسط تابع هزینه شبکه معکوس ترکیب می‌شود. رسیدن به نقطه بهینه و پایدار در شبکه‌های تقابلی مولد، دشوار است. برای بهینه‌شدن وضعیت پاسخ، آرژوفسکی<sup>۱۰</sup> و همکاران در [۳۲] روش بهتری ارائه کردند که یادگیری را راحت‌تر و از رخداد فروپاشی حالت<sup>۱۱</sup> در شبکه تقابلی مولد جلوگیری می‌کند. ژائو و همکاران باتوجه به بهبودهای متنوعی که پژوهش‌های مختلف، نظیر [۳۲] و [۳۴]–[۳۶] در یادگیری شبکه تقابلی مولد ایجاد کردند، تابع هزینه‌ای به صورت رابطه (۲۱) ارائه کردند:

$$\min_{\theta} \max_{\omega} \mathbb{E}_{x \sim p_x} [C_{\omega}(x)] - \lambda \cdot \mathbb{E}_{z \sim p_z} [C_{\omega}(G_{\theta}(z))] \quad (21)$$

در رابطه (۲۱) منظور از  $C_{\omega}$  تابع منتقد<sup>۱۲</sup> است. همچنین، از مقدارهای ۰.۱ و ۱ به ترتیب، برای حوزه تصویر و متن جهت تخصیص به شاخص  $\lambda$  در رابطه (۲۱) استفاده کردند.

ژائو و همکاران به منظور تولید نمونه خصمانه، نخست یک شبکه WGAN را روی مجموعه داده اصلی  $X$  آموزش می‌دهند. شبکه یادشده شامل شبکه مولد  $G_{\theta}$  بوده و وظیفه نگاشت بازنمایی بردار  $z \in \mathbb{R}^d$  به نمونه‌های تولیدشده نظیر  $x$  از دامنه  $X$  را داراست. آن‌ها شبکه معکوس  $J$  که وظیفه نگاشت نمونه‌های داده به بازنمایی  $z$  را داراست، جداگانه آموزش می‌دهند. در نهایت، در روند شبکه تقابلی مولد، اختلاف بازنمایی تولیدشده با خروجی

حمله جعبه سیاه با استفاده از بهینه‌سازی مرتبه صفر: چن<sup>۱</sup> و همکاران در [۲۹] رویکردی مستقل از گرادیان ارائه کردند. آن‌ها برخلاف بیشتر روش‌های تولید نمونه خصمانه که اغلب با گرادیان کار می‌کنند، روشی ارائه کردند که نیازی به استفاده از گرادیان نداشته باشد. در نتیجه، روش آن‌ها گزینه 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  $f$  مطرح‌شده در پژوهش کارلینی و وگنر [۳۰] به رابطه (۱۹) دست یافتند.

$$f(\hat{x}) = \max(\max\{\log[Z(\hat{x})]_i; i \neq t\} - \log[Z(\hat{x})]_{y_{target}}, -\kappa) \quad (19)$$

گفتنی است که آن‌ها با استفاده از مشتق متقارن، تخمینی از گرادیان نیز به صورت رابطه (۲۰) به دست آورده‌اند:

$$\frac{\partial f(x)}{\partial x_i} \approx \frac{f(x + he_i) - f(x - he_i)}{2h} \quad (20)$$

منظور از  $h$  در رابطه (۲۰)، عددی ثابت با مقدار پیشنهادی ۰.۰۰۰۱ است.  $e_i$  نیز بردار پایه استاندارد است که فقط درایه  $i$ امین عنصر آن یک است. تخمین خطای آن‌ها در مرتبه  $O(h^2)$  است. از آنجاکه در روش FGSM تابع علامت گرادیان استفاده می‌شود، چن و همکاران بر این عقیده بودند که نیازی به در اختیار داشتن مقدار دقیق گرادیان نبوده است؛ بنابراین، نظر مطرح‌شده آن‌ها نیز بر همین اصل بنا نهاده شده که با استفاده از تقریب و تابع واسط روال کار پیش برده شود. گفتنی است آن‌ها به منظور ارزیابی بهتر، تخمین گرادیان را چندین بار تکرار می‌کنند. تخمین هسین<sup>۲</sup> مطرح‌شده در پژوهش آن‌ها نیز به صورت رابطه (۲۲) است.

$$\frac{\partial^2 f(x)}{\partial x_{ii}^2} \approx \frac{f(x + he_i) - 2f(x) + f(x - he_i)}{h^2} \quad (22)$$

آن‌ها از نام zoo<sup>۳</sup> 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 محاسبه تخمین گرادیان امری زمان‌بر است. نتایج نشان داده است که روش آن‌ها از روش کارلینی و وگنر [۱۶] نیز بهتر عمل می‌کند.

<sup>4</sup> Generating Natural Adversarial Example

<sup>5</sup> Zhao

<sup>6</sup> Generative Adversarial Network

<sup>7</sup> Wasserstein

<sup>8</sup> Wasserstein Generative Adversarial Network

<sup>۹</sup> منظور از معکوس بودن شبکه این است که از فضای داده به فضای نهان قابلیت نگاشت داده داشته باشد.

<sup>۱۰</sup> Arjovsky

<sup>۱۱</sup> Mode collapse

<sup>۱۲</sup> Critic

<sup>۱</sup> Chen

<sup>۲</sup> Hessian

<sup>۳</sup> Zero Order Optimization

### ۳- دفاع<sup>۴</sup> در برابر نمونه خصمانه

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

#### تعریف دفاع

به فرآیندهای مقابله با آسیب‌پذیری شبکه‌های عصبی عمیق، نسبت به نمونه‌های خصمانه دفاع گفته می‌شود. این فرآیندها شامل روش‌های متنوعی هستند که پاسخ الگو را نسبت به ورودی خصمانه کنترل می‌کنند. همچنین، در برخی روش‌ها دقت و تعمیم‌پذیری الگو نیز افزایش می‌یابد. (جدول ۱) طبقه‌بندی روش‌های حمله پرداخته شده در این مقاله. T بیانگر هدفمند بودن رویکرد مدنظر، W و B به ترتیب روش‌های جعبه سفید و جعبه سیاه 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            | هدفمند (T) | نوع حمله (W/B) | تعداد پارامتر | نوع 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 (شکل ۸)، روش‌های دفاع و مقابله با نمونه‌های خصمانه به دو دسته روش‌های کنشی<sup>۵</sup> و روش‌های پیش‌کنشی<sup>۶</sup> تقسیم‌بندی می‌شوند [۳۷]. روش‌های کنشی پس از ساخت شبکه عصبی (الگو)، درصدد بررسی آسیب‌پذیری و مقابله با نمونه خصمانه برمی‌آییم. در برخی روش‌ها پس از تشخیص خصمانه بودن نمونه ورودی، آن ورودی از ادامه فرآیند

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

$$\begin{aligned} \hat{x} &= G_{\theta}(\hat{z}) \\ \hat{z} &= \underset{\tilde{z}}{\operatorname{argmin}} \|\tilde{z} - \mathcal{I}_V(x)\| \\ s.t. \quad & F(G_{\theta}(\tilde{z})) \neq F(x) \end{aligned} \quad (23)$$

حمله تک پیکسل<sup>۱</sup>: سو<sup>۲</sup> و همکاران در [۲۷] تلاش

کردند تا با تغییر دادن یک پیکسل، نمونه خصمانه تولید نمایند. آن‌ها تصویر ورودی را به صورت برداری در نظر گرفتند که هر عنصر آن نشان‌دهنده یک پیکسل از تصویر است. مسئله بهینه‌سازی پیدا کردن اختلال در نگاه آنان به صورت رابطه (۲۴) بیان می‌شود:

$$\begin{aligned} \min_{\eta} \quad & F(x + \eta) \\ s.t. \quad & \|\eta\|_0 \leq d \end{aligned} \quad (24)$$

در رابطه (۲۴) منظور از ثابت d، عددی کوچک و این عدد برای حمله تک پیکسل عدد یک است. درواقع، منظور این است که تنها در d بعد می‌توان تغییرات ایجاد کرد و در دیگر ابعاد تغییری ایجاد نمی‌شود. آن‌ها به منظور پیدا کردن پاسخ بهینه از تکامل تفاضلی<sup>۳</sup> استفاده کردند. برتری استفاده از تکامل تفاضلی عدم نیاز به گرادیان شبکه 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 نسبت به نمونه‌های خصمانه آسیب‌پذیر هستند و نیاز است بر مقاومتی آن‌ها تمرکز شود. در ادامه، انواع روش‌های مقابله با نمونه‌های خصمانه و رویکردهای دفاع در مقابل آن‌ها را مرور خواهیم کرد.

#### تعریف دفاع

به فرآیندهای مقابله با آسیب‌پذیری شبکه‌های عصبی عمیق، نسبت به نمونه‌های خصمانه دفاع گفته می‌شود. این فرآیندها شامل روش‌های متنوعی هستند که پاسخ الگو را نسبت به ورودی خصمانه کنترل می‌کنند. همچنین، در برخی روش‌ها دقت و تعمیم‌پذیری الگو نیز افزایش می‌یابد. (جدول ۱) مقایسه روش‌های مختلف تولید نمونه خصمانه انجام شده است.

<sup>۴</sup> Defense

<sup>۵</sup> Reactive

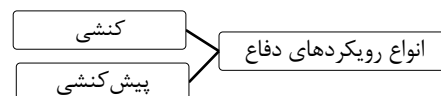
<sup>۶</sup> Preactive

<sup>۱</sup> One-pixel attack

<sup>۲</sup> Su

<sup>۳</sup> Differential evolution

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

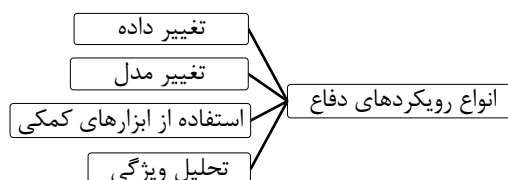


(شکل ۸) - انواع رویکردهای دفاع بر اساس نحوه واکنش

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 دسته تغییر داده<sup>۱</sup>، تغییر الگو<sup>۲</sup> و استفاده از ابزارهای کمکی<sup>۳</sup> و تحلیل فضای ویژگی تقسیم‌بندی کرد. همچنین، در برخی پژوهش‌های اخیر نظیر [۲۴] و [۴۲] - [۴۴] به میزان درهم‌تنیدگی فضای ویژگی یادگرفته شده نیز اشاره شده است. 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 خواهیم داشت:

### تغییر داده

در این روش به منظور 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 مرحله آموزش تنوع بیشتری از داده را مشاهده کرده و در نتیجه، 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<sup>۴</sup>: در این روش نمونه‌های خصمانه به مجموعه دادگان آموزشی اضافه می‌شوند. به بیان دیگر نمونه‌های خصمانه از این پس به عنوان نمونه‌های مجاز<sup>۵</sup> محسوب و در کنار مجموعه دادگان آموزشی در آموزش استفاده می‌شوند. سپس، الگو با مجموعه دادگان جدید آموزش می‌بیند و همین امر موجب می‌شود تا نسبت به حالت پیشین، نسبت به حملات خصمانه مقاوم‌تر عمل کند. البته این موضوع جای بحث دارد و نمی‌توان بیان کرد که نسبت به همه نمونه‌های خصمانه مقاوم خواهد بود. زیرا نمونه‌های خصمانه با روش‌های متنوعی تولید می‌شوند. همان‌طور که بیان شد، در پژوهش‌های گوناگونی از آموزش خصمانه به اضافه کردن نمونه‌های خصمانه<sup>۶</sup> به مجموعه دادگان اصلی و افزایش مجموعه دادگان و آموزش دوباره الگو روی آن‌ها یاد کرده‌اند.

سژدی و همکاران در [۹] به اضافه کردن نمونه‌های خصمانه و در نتیجه، افزایش مقاوم‌سازی الگو نسبت به نمونه‌های خصمانه اشاره کردند. این بحث توسط گودفلو و همکاران در مقاله [۸] بسط داده شده و جوانب مختلفی از آن پوشش داده شد. آن‌ها از آن به نام تابع هزینه خصمانه<sup>۷</sup> یاد و بیان کردند که می‌توان با اضافه کردن یک عبارت منظم‌ساز<sup>۸</sup> به تابع هزینه، تأثیری به نسبت برابر با افزودن نمونه‌های خصمانه به مجموعه داده ایجاد کرد. امروزه اصطلاح «آموزش خصمانه» اغلب به اضافه کردن این عبارت منظم‌ساز مبتنی بر حمله به تابع هزینه اصلی اطلاق می‌شود. برای مثال، آن‌ها تابع هزینه خصمانه را برای روش حمله FGSM به صورت ترکیب خطی از تابع هزینه و عبارت منظم‌ساز همانند رابطه (۲۵) بیان کرده‌اند. به عبارت منظم‌ساز اضافه شده، عبارت خصمانه<sup>۹</sup> نیز گفته می‌شود.

$$J^*(\theta, x, y) = \alpha J(\theta, x, y) + (1 - \alpha) J(\theta, x + \epsilon \text{sign}(\nabla_x J(\theta, x, y)), y) \quad (25)$$

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

<sup>4</sup> Adversarial training

<sup>5</sup> Legalized

<sup>6</sup> نمونه‌های خصمانه هر مرحله از یادگیری مدل

<sup>7</sup> Adversarial cost function

<sup>8</sup> Regularization term

<sup>9</sup> Adversarial term

<sup>1</sup> Modifying data

<sup>2</sup> Modifying model

<sup>3</sup> Auxiliary tools

منظم‌ساز به‌عنوان عبارت جری‌مه‌کننده لحاظ شود و با تابع هزینه اصلی جمع شود. در نهایت، وضعیت بهینه زمانی خواهد بود که هر دو عبارت دخیل در تابع هزینه کلی کمینه شوند (تابع هزینه قبلی و عبارت منظم‌ساز جدید). عبارت منظم‌ساز آموزش خصمانه رویکرد سژدی و همکاران، مبتنی بر حمله FGSM است. استفاده از حمله PGD در آموزش خصمانه توسط مادری و همکاران در [۲۴] پیشنهاد شده و دارای هزینه محاسباتی بالایی نسبت به FGSM است. ونگ<sup>۱</sup> و همکاران در [۴۱] با افزودن یک مرحله اولیه (مرحله مقداردهی تصادفی اختلال)، عملکرد استفاده از FGSM در آموزش خصمانه را تا اندازه استفاده از PGD در آموزش خصمانه افزایش دادند و آن را FFGSM<sup>۲</sup> نامیدند. این امر موجب افزایش سرعت آموزش و کاهش حجم محاسبات شده است. 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 نسبت به FGSM نیز استفاده شده است.

**آموزش خصمانه جمعی:**<sup>۳</sup> ترامر<sup>۴</sup> و همکاران در [۴۲]، نظر آموزش خصمانه جمعی را مطرح کردند. آموزش خصمانه جمعی، طیف بیشتری از نمونه‌های خصمانه را پوشش می‌دهد و روال عملکرد آن توسعه آموزش خصمانه است. این امر موجب شده تا نسبت به آموزش خصمانه عادی بهتر عمل کند و نسبت به نمونه‌های خصمانه بیشتری مقاوم باشد. آن‌ها بیان کردند که آموزش خصمانه به کمینه سراسری نزدیک می‌شود و در نتیجه، نسبت به حملات جعبه سیاه الزاماً مقاوم نیست. درواقع، آن‌ها اصل انتقال‌پذیری را که گاهی در حملات جعبه سیاه 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 خصمانه جمعی را ارائه کردند. همچنین، بیان کردند که آموزش خصمانه جمعی به تولید اختلال‌های متنوع منجر شده و در نتیجه، افزایش تنوع داده‌های آموزشی را به همراه دارد؛ بنابراین، منطقی است که نسبت به نمونه‌های خصمانه 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

خصمانه جمعی، به‌گونه‌ای نسخه توسعه‌یافته آموزش خصمانه است؛ بنابراین، از طریق الگوهای از پیش تعلیم‌دیده شده موجود، نمونه خصمانه تولید و در نتیجه، موجب تکمیل مجموعه داده خصمانه 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 نسبت به نمونه‌های خصمانه بیشتری مقاوم باشد.

**پنهان‌سازی گرادیان:**<sup>۵</sup> ترامر و همکاران در پژوهش [۴۲] نیز رویکردی جهت مقابله با روش‌های حمله مبتنی بر گرادیان ارائه کردند. این رویکرد دارای قابلیت دفاع در مقابل روش‌های حمله مبتنی بر گرادیان نظیر FGSM و نسخه‌های توسعه‌یافته آن است. آن‌ها بیان کردند که می‌توان الگوهای 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 مبتنی بر گرادیان به راحتی مقابله می‌کنند.

پیرو و همکاران در [۴۳] رویکردی ارائه کردند که موجب نقض روش دفاع «پنهان‌سازی گرادیان» شد. درواقع، آن‌ها پیشنهاد تعلیم یک دسته‌بند جایگزین<sup>۶</sup> با نتایج مشابه را دادند. سپس، با استفاده از گرادیان الگوی جایگزین، تولید نمونه 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:**<sup>۷</sup> خاصیت انتقال‌پذیری جزء ویژگی‌هایی است که با تغییر ماهیت شبکه‌ها همچنان در برخی موارد مشاهده می‌شود. با توجه به این که از انتقال‌پذیری نمونه‌های خصمانه در حملات جعبه سیاه به‌وفور استفاده می‌شود، حسینی و همکاران در [۴۴] رویکردی جهت کاهش انتقال‌پذیری معرفی کردند. آن‌ها با معرفی روش سه مرحله‌ای برچسب‌گذاری پوچ<sup>۸</sup>، موفق به کاهش انتقال‌پذیری از یک شبکه به شبکه دیگر شدند. نظر آنها اضافه‌شدن برچسب پوچ به مجموعه برچسب‌های خروجی شبکه است. روش آن‌ها با دیدن نمونه خصمانه به جای افزایش احتمال برچسب‌های اصلی، احتمال

<sup>۵</sup> Gradient hiding

<sup>۶</sup> Surrogate

<sup>۷</sup> Blocking the transferability

<sup>۸</sup> Null labeling

<sup>۱</sup> Wang

<sup>۲</sup> Fast adversarial training using FGSM (FFGSM)

<sup>۳</sup> Ensemble adversarial training

<sup>۴</sup> Tramèr

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

### فشرده‌سازی داده<sup>۱</sup>: زیوگیت<sup>۲</sup> 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 مبتنی بر گرادینان نظیر **FGSM** تأثیر به‌سزایی دارد. دس<sup>۳</sup> 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 مبتنی بر گرادینان 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 و گنر [۱۶] چندان مؤثر نیستند.

اختر و همکاران در [۴۷] به ارائه الگویی جهت دفاع نسبت به حملات خصمانه فراگیر پرداختند. در روش آن‌ها یک مرحله قبل از ارسال داده به الگو اضافه شده‌است. در مرحله تعبیه‌شده آن‌ها، فرایند اصلاح<sup>۴</sup> داده انجام می‌شود. در روش آن‌ها از تشخیص‌دهنده اختلال استفاده شده که از تبدیل کسینوسی گسسته<sup>۵</sup> که اغلب در فشرده‌سازی نیز استفاده می‌شود، بهره گرفته شده‌است. مشکل عمده رویکردهای دفاع مبتنی بر 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<sup>۶</sup>: شی<sup>۷</sup> و همکاران در [۴۸]** تلاش کردند تا تأثیر تغییر اندازه داده خصمانه در دسته‌بند را بررسی کنند. آن‌ها نشان دادند که تغییر اندازه تصادفی موجب کاهش تأثیرگذاری نمونه خصمانه بر الگو می‌شود. همچنین، بیان کردند که قراردادن بافت<sup>۸</sup> تصادفی در تصاویر موجب کاهش تأثیرگذاری نمونه‌های خصمانه می‌شود. آن‌ها با تعریف یک ماژول<sup>۹</sup> در راستای افزودن یک مرحله پیش‌پردازش غیرمشق‌پذیر به نام تعبیه

محلی خطی<sup>۱۰</sup> توانستند بر اختلال موجود در نمونه‌های خصمانه غالب شوند. گفتنی است که برای جامعیت بهتر، آن‌ها در مرحله آموزش نیز از آن ماژول استفاده کرده‌اند. همچنین، پس از مرحله عبور از ماژول، نوفه تصادفی گوسی نیز به آن افزوده می‌شود که این امر بهبوددهنده مقاومت الگوست.

**تفکیک ویژگی‌های مقاوم از غیرمقاوم: الیاس<sup>۱۱</sup> و همکاران در [۴۹]**، نگرش دیگری به مسئله 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 مجموعه 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، این که موهای بدن یک گربه الگوی افقی دارد و چندان 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 مؤثر در تشخیص آنها هستند (که شبکه‌های عصبی بعضاً از آن‌ها غافلند). آن‌ها با تفکیک ویژگی‌های داده به سه دسته ویژگی‌های  $\rho$ -مفید، ویژگی‌های  $\gamma$ -مقاوم و ویژگی‌های مفید غیرمقاوم، سعی کردند مسئله را با بیانی ریاضی بازنویسی نمایند. زمانی یک ویژگی  $\rho$ -مفید خواهد بود که همبستگی بین آن ویژگی‌ها، با برچسب نهایی در بالاترین سطح ممکن قرارگیرد. درواقع، اگر مجموعه دادگان از توزیع داده شده  $D$  نمونه‌برداری شوند، ویژگی‌های  $\rho$ -مفید به‌صورت رابطه (۲۶) قابل بیان هستند:

$$\mathbb{E}_{(x,y) \sim D} [y \cdot f(x)] \geq \rho \quad (26)$$

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

$$\mathbb{E}_{(x,y) \sim D} \left[ \inf_{\delta \in \eta} y \cdot f(x + \delta) \right] \geq \gamma \quad (27)$$

<sup>10</sup> Linear Locally Embedding (LLE)

<sup>11</sup> Ilyas

<sup>1</sup> Data compression

<sup>2</sup> Dziugaite

<sup>3</sup> Das

<sup>4</sup> Rectifying

<sup>5</sup> Discrete Cosine Transform

<sup>6</sup> Data randomization

<sup>7</sup> Xie

<sup>8</sup> Texture

<sup>9</sup> Module

همان‌طور که در رابطه (۲۷) نیز آورده شده، کمترین اختلالی که زیر مجموعه اختلال‌های خصمانه باشد و موجب افزایش همبستگی ویژگی‌های ورودی و برچسب شود، و از حد آستانه  $\gamma$  بیشتر باشد، از ویژگی‌های  $\gamma$ -مقاوم به شمار می‌رود. ویژگی‌های مفید و غیرمقاوم، ویژگی‌های  $\rho$ -مفیدی هستند که به‌ازای هیچ مقدار  $\gamma \geq 0$ ،  $\gamma$ -مقاوم نیستند. این ویژگی‌ها موجب افزایش صحت دسته‌بندی در شرایط غیرخصمانه هستند، اما زمانی که 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 رابطه (۲۸) استخراج نمایند.

$$\min_{x_r} \|g(x_r) - g(x)\|_2 \quad (28)$$

در رابطه (۲۸) منظور از  $g$ ، یک بازنمایی از ورودی و منظور از  $x_r$ ، ویژگی‌های مقاوم متناظر با هر نمونه‌ی  $x$  است. آن‌ها روند پیدا کردن نقطه بهینه مسئله (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 داده جدیدی تولید می‌کنند که شامل ویژگی‌های مقاوم است. سپس، آموزش شبکه روی مجموعه داده 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 نشده است. این ضعف در پژوهش‌هایی نظیر **MagNet** [۵۰] نیز بیان شده که توسط طیف مختلفی از پژوهشگران نقد شده است.

### تغییر الگو

به‌منظور دفاع در مقابل حملات خصمانه می‌توان تغییراتی در الگوی یادگیری به‌وجود آورد 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<sup>۱</sup>، خالص‌سازی دفاعی<sup>۲</sup>، فشردن ویژگی<sup>۳</sup>، دفاع پوشش<sup>۴</sup> و شبکه انقباضی عمیق<sup>۵</sup>.

**عبارت منظم‌ساز:** عبارت منظم‌ساز اغلب با عنوان عبارت‌های جریمه<sup>۶</sup> در تابع هزینه ظاهر می‌شوند. این عبارات قابلیت کنترل بخش‌های دیگری از هدف مدنظر را به حل مسئله بهینه‌سازی اضافه می‌کنند. بیگیو<sup>۷</sup> و همکاران در [۵۱] تلاش کردند تا با استفاده از اضافه کردن عبارت منظم‌ساز، ماتریس کرنل ایجادشده در مسئله بهینه‌سازی **SVM**<sup>۸</sup> را نسبت به نمونه خصمانه بهینه کنند.

**خالص‌سازی دفاعی:** پپرنو و همکاران در [۲۸] روشی جهت دفاع در مقابل نمونه‌های خصمانه، مبتنی بر آموزش خالص‌سازی دانش<sup>۹</sup> [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 شده است. همان‌طور که می‌دانیم، برچسب تخمینی شبکه‌های عصبی در حالت عادی با دریافت لاجیت و ارسال آن به لایه پیشین نرم تولید می‌شود. فرآیند خالص‌سازی دانش توسط شاخصی به نام دما (که با  $T$  نشان می‌دهند) انجام می‌شود. در حالت عادی مقدار این شاخص یک ( $T = 1$ )، در نظر گرفته می‌شود. هرچه میزان  $T$  بیشتر باشد، خروجی توزیع نرم‌تری بین رده‌ها خواهد داشت.

همان‌طور که در رابطه (۲۹) بیان شد، در یادگیری خالص‌سازی دانش، از شاخص نرم‌کننده‌ای به نام دما ( $T$ ) استفاده می‌شود. حداقل مقدار شاخص  $T$  مقدار یک است و هرچه این میزان افزایش یابد، طبق رابطه (۲۹)، موجب نرم شدن خروجی شبکه می‌شود. بنابراین، خروجی

$$O_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \quad (29)$$

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

<sup>1</sup> Regularization term

<sup>2</sup> Defensive distillation

<sup>3</sup> Feature squeezing

<sup>4</sup> Mask defense

<sup>5</sup> Deep Contractive Network (DCN)

<sup>6</sup> Penalty terms

<sup>7</sup> Biggio

<sup>8</sup> Supported Vector Machine

<sup>9</sup> Knowledge distillation

پیشنهادی خود با مقایسه خروجی الگو به‌ازای داده ورودی و داده فشرده شده با یک حد آستانه مشخص، بیان کردند که نمونه خصمانه از نمونه تمیز قابل شناسایی 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 فشرده کردن ویژگی در داده تصویری پیشنهاد کردند. کاهش عمق رنگی هر پیکسل در تصویر و استفاده از نرم‌کننده مکانی برای کاهش تفاوت بین پیکسل‌های مستقل، از جمله این پیشنهادها هستند. آن‌ها با انجام آزمایش‌های مختلف نشان دادند که چارچوب فشرده کردن ویژگی، عامل مؤثری در افزایش مقاومت الگوست. همچنین، نکته 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.

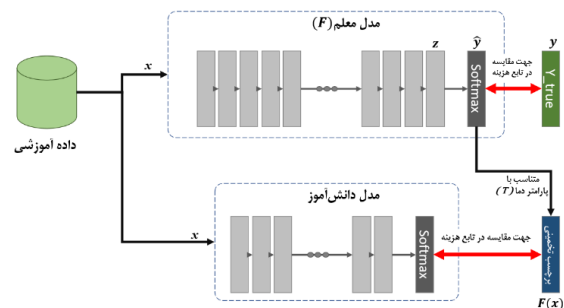
**دفاع پوشش: گنو<sup>۳</sup> و همکاران در [۵۵]** سعی کردند روشی جهت شناسایی ویژگی‌های مؤثر جهت مقاوم کردن الگو نسبت به حملات ارائه کنند. آن‌ها با اضافه کردن یک لایه پوشش قبل از لایه‌های تمام‌متصل، ویژگی‌های مؤثر را برای مرحله دسته‌بندی انتخاب و مبتنی بر میزان فاصله خروجی بردار ویژگی هرلایه در زمان ورود داده تمیز و داده خصمانه، اقدام به پوشش ویژگی‌های مدنظر می‌کنند. **شبکه عمیق انقباضی: گو<sup>۴</sup> و همکاران در [۵۶]** تلاش‌های متعددی در بررسی میزان مقاوم بودن انواع شبکه‌های خودرمزگذار انجام دادند. آن‌ها با روش‌های متعددی نظیر افزودن نوفه به داده خصمانه و استفاده از خودرمزگذار حذف‌کننده نوفه<sup>۵</sup> تلاش کردند تأثیر خصمانه را کاهش دهند. ولی متوجه شدند که با این نوع آموزش، دوباره شبکه قابلیت نفوذ خواهد داشت. 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 رابطه تابع هزینه آن به صورت رابطه (۳۰) است:

$$J(\theta) = \sum_{i=1}^m (L(x^{(i)}, y^{(i)}) + \lambda \|\nabla_{x^{(i)}} h(x^{(i)})\|_F^2) \quad (30)$$

<sup>۳</sup> Gao

<sup>۴</sup> Gu

<sup>۵</sup> Denoising Auto Encoder (DAE)

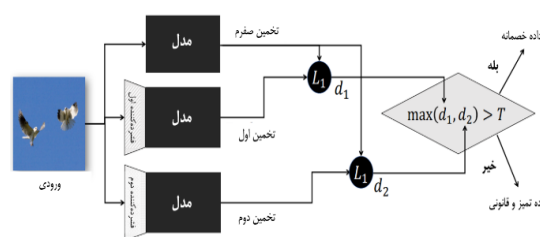


(شکل - ۱۰) - مثال انتقال دانش از الگوی معلم به الگوی

دانش آموز در نظر پژوهش خالص سازی 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، شبکه اولیه (شبکه معلم) با دمای مشخص  $T$  در بیشینه نرم آموزش می‌بیند. سپس، با دادن مجموعه دادگان آموزشی  $X$  به شبکه اولیه  $F$  و دریافت خروجی آن، بردار احتمال شبکه اولیه به دست می‌آید. سپس، از آن بردار احتمال به جای برچسب مجموعه دادگان اولیه جهت آموزش همان ساختار شبکه در دمای مشخص  $T$  استفاده می‌شود. این امر موجب می‌شود تا دانش کسب شده از احتمال رده‌های مختلف به ازای ویژگی‌های داده ورودی بهتر بازنمایی شود. در نتیجه، به جای استفاده از بردارهای تک نقطه<sup>۱</sup>، که هیچ ارتباط مفهومی را بین رده‌های مختلف نشان نمی‌دهد، از بردارهای مفهوم دار پیش‌بینی شبکه  $F$  استفاده می‌شود. گفتنی است که موثر بودن این روش بسته به شرایط خاص آموزش آن دارد. درواقع، اگر متخاصم بداند که در فرآیند آموزشی شبکه از فرآیند خالص سازی استفاده شده است، می‌تواند با رویکرد دیگری به آن حمله کند که در پژوهش کارلینی و وگنر [۵۳] به آن اشاره شده است.



(شکل - ۱۱) - چارچوب تشخیص نمونه خصمانه روش

فشرده‌سازی ویژگی شو و همکاران [۵۴]

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

**فشرده‌سازی ویژگی: شو<sup>۲</sup> و همکاران در [۵۴]** فضای قابل جستجوی متخاصم را به نحوی کاهش داده‌اند. همان‌طور که در (شکل - ۱۱) نمایان شده، آن‌ها در رویکرد

<sup>۱</sup> تنها یکی از عناصر آن بردار مقدار یک است.

<sup>۲</sup> Xu

که در آن منظور از  $m$  تعداد نمونه‌های آموزشی،  $L$  تابع هزینه عادی شبکه خودرمزگذار،  $\lambda$  شاخص عبارت منظم‌ساز و  $\| \cdot \|_F$  نرم فروبنیوس<sup>۱</sup> است. هدف کلی تابع هزینه رابطه (۳۰) ذخیره‌کردن اطلاعات مفید از داده ورودی و کاهش توجه به ویژگی‌های مختل شده 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 رابطه (۳۱) ارائه دهند. گفتنی است آن‌ها، همچنین، اشاره کردند که تابع هزینه رابطه (۳۱) بهینگی سراسری را تضمین نمی‌کند، بلکه موجب محدودکردن ظرفیت شبکه‌عصبی می‌شود.

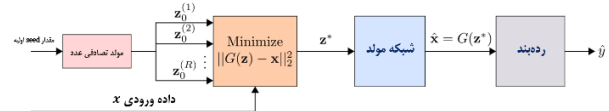
$$J(\theta) = \sum_{i=1}^m \left( L(t^{(i)}, y^{(i)}) + \sum_{j=1}^{H+1} \lambda_j \left\| \frac{\partial h_j^{(i)}}{\partial h_{j-1}^{(i)}} \right\|_2 \right) \quad (31)$$

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

### استفاده از ابزارهای کمکی

بعضی از روش‌های مبتنی بر استفاده از ابزارهای جانبی عبارتند از: دفاع مبتنی بر شبکه تقابلی مولد، شبکه مگنت<sup>۲</sup> و حذف‌کننده نوفه هدایت‌شده بازنمایی سطح بالا<sup>۳</sup>. در ادامه، این روش‌ها معرفی می‌شوند.

**دفاع مبتنی بر شبکه تقابلی مولد:** سمنقویی و همکاران در [۵۷] سازوکاری به نام DefenseGAN جهت دفاع نسبت به حملات جعبه سیاه، جعبه سفید و کاهش تأثیرگذاری اختلال خصمانه بر الگو ارائه کردند. روش آن‌ها مبتنی بر 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]

همان‌طور که در (شکل ۱۲) بیان شده، نظر اصلی این است که داده ورودی به فضای یادگرفته‌شده شبکه مولد نگاشت داده می‌شود. درواقع، نزدیک‌ترین داده قابل تولید توسط مولد به عنوان داده مفید به شبکه اصلی داده می‌شود. مشکل اصلی این روش این است که نخست باید یک شبکه تقابلی مولد آموزش دیده 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 چالش برانگیز است. همچنین، در مرحله اول نیاز است تا بهترین  $z$  ممکن برای مرحله بعد انتخاب شود. زمان‌بر بودن مرحله شناسایی  $z$  بهینه نیز از معضلات دیگر این روش به شمار می‌رود. همچنین، در صورتی که به شبکه تقابلی مولد از قبل حمله شده باشد، این بحث به کلی منتفی می‌شود. بنابراین، پژوهش آن‌ها در شرایط بسیار خاصی قابلیت استفاده و اجرا دارد.

روش آن‌ها نیازمند انتخاب بهترین  $z$  ( $z^*$ ) و بررسی آن با حد آستانه  $\rho$  موجود در رابطه (۳۲) انجام می‌پذیرد. پیدا کردن مقدار آستانه بهینه نیز از چالش‌های دیگر پژوهش آن‌هاست.

دارای اختلال خصمانه

$$\|G(z^*) - x\|_2^2 \leq \rho \quad (32)$$

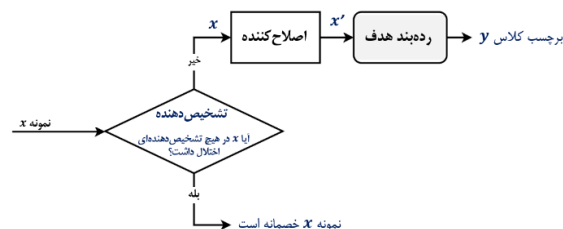
داده تمیز

**شبکه مگنت:** مگ<sup>۴</sup> و چن<sup>۵</sup> در [۵۰] چارچوبی جهت حفاظت از شبکه‌های عصبی نسبت به نمونه‌های خصمانه ارائه کردند. طبق (شکل ۱۳) در نظر پیشنهادی آن‌ها چند شبکه تشخیص‌دهنده<sup>۶</sup> و یک شبکه اصلاح‌کننده<sup>۷</sup> وجود دارد. شبکه‌های تشخیص‌دهنده اختلاف نمونه‌های عادی و خصمانه را مقایسه می‌کنند. شبکه اصلاح‌کننده نیز داده‌های خصمانه را به منی‌فولد نمونه‌های عادی بازگشت می‌دهد.

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

<sup>4</sup> Meng  
<sup>5</sup> Chen  
<sup>6</sup> Detector  
<sup>7</sup> Reformer

<sup>1</sup> Frobenius  
<sup>2</sup> MagNet  
<sup>3</sup> High-level representation guided denoiser



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

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

### حذف‌کننده نوفه هدایت‌شده بازنمایی سطح بالا:

اغلب حذف‌کننده‌های نوفه استاندارد، از مشکل تشدید خطا رنج می‌برند (نظیر آن‌هایی که در سطح پیکسل کار می‌کنند). به عبارت دیگر، نوفه باقیمانده<sup>۱</sup> کوچک، به تدریج تقویت شده و موجب خطای دسته‌بندی می‌شود. لیاو<sup>۲</sup> و همکاران در [۵۹] تلاش کردند که این مشکل را با ارائه تابع هزینه جدید با ترکیب اختلاف خروجی الگو روی

$$D(T(x), T(\hat{x})) \ll \varepsilon \quad (35)$$

تصویر تمیز و تصویر مختل شده برطرف کنند. مطابق رابطه (۳۳)، آن‌ها با استخراج نرم اول خروجی بازنمایی لایه  $l$  ام، برای نمونه داده تمیز و مختل شده به سیگنالی دست پیدا کرده‌اند که از آن برای ادامه مسیر استفاده کردند.

$$L = \|f_l(\hat{x}) - f_l(x)\| \quad (33)$$

در پژوهش یوسا<sup>۳</sup> 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<sup>۴</sup> 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 مبتنی بر رمزگذاری تنک کانولوشنال، فضایی به‌طور تقریبی طبیعی کم‌بعدی را ایجاد و بیان کردند که نمونه‌های خصمانه به‌ندرت در آن وجود دارند. همچنین، آن‌ها با ارائه لایه تبدیل تنک بین تصویر ورودی و اولین لایه شبکه نیز موجب بهینه کردن 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، فاصله زیادی‌تری نسبت به 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 حذف‌کننده نوفه از پیش‌تعلیم‌دیده، خوشه مربوط به داده ورودی را شناسایی می‌کنند. سپس، خوشه انتخابی و نقشه‌های ویژگی تنک (همراه با فیلترهایشان) را به‌صورت فرهنگت<sup>۵</sup> 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 داده طبیعی تبدیل‌یافته  $(T(\hat{x}))$  و داده خصمانه تبدیل‌یافته  $(T(x))$  را برای داده‌های یکسان کمینه کنند (در حد عدد ثابت کوچک). (E)

در رابطه (۳۵) منظور از  $D$  فاصله و  $T$  داده ورودی به فضای طبیعی داده است. آن‌ها برای فراگرفتن فرهنگ لغت مناسب در ساختار کانولوشنی، از حل مسئله بهینه‌سازی رابطه (۳۴) بهره بردند:

در رابطه (۳۴) منظور از  $\otimes$  عملیات کانولوشن،  $C$  تعداد کانال‌های ورودی،  $K$  تعداد فیلترهای هر کانال ورودی،  $f_{i,c}$  مجموعه‌ای از فیلترها و  $z_i$  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 حمله مرسوم نظیر FGSM، BIM، DeepFool و C&W عملکرد خوبی داشته است.

$$\begin{aligned} \min & \frac{1}{2} \sum_{c=1}^C \|x_c - T(x)_c\|_2^2 + \lambda \sum_{i=1}^K \|z_i\|_1 \\ \text{s.t.} & T(x)_c = \sum_{i=1}^K f_{i,c} \otimes z_i \\ \text{s.t.} & \|f_{i,c}\|_2^2 = 1 \\ & 1 \leq i \leq K, 1 \leq c \leq C \end{aligned} \quad (34)$$

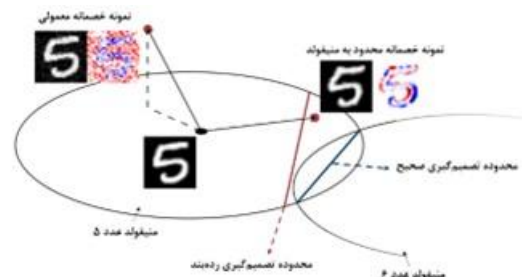
## ۴- بحث در مورد تحلیل درهم‌تنیدگی ویژگی‌ها و تأثیر آن در تعمیم‌پذیری و مقاومت‌سازی

در این قسمت به مرور برخی پژوهش‌ها در خصوص تحلیل ویژگی‌های فراگرفته‌شده در الگوهای یادگیری ماشین با تأکید بر مسئله تحلیل حساسیت نسبت به نمونه‌های خصمانه پرداخته شده‌است. سیپراس<sup>۶</sup> و

<sup>5</sup> Dictionary  
<sup>6</sup> Tsipras

<sup>1</sup> Residual  
<sup>2</sup> Liao  
<sup>3</sup> Uesato  
<sup>4</sup> Sun

همکاران در [۶۲] به وجود یک تنش بین تعمیم و مقاومت‌سازی 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 مقاومت‌سازی الگو مصالحه ایجاد کرد و بررسی سطحی آن‌ها موجب رسیدن به پاسخ موجهی نمی‌شود. همچنین، آن‌ها اشاره کردند که مقاومت‌سازی در نتیجه آموزش عادی پدید نمی‌آید و نیاز است با به‌کارگیری یک رویکرد بهینه انجام شود.

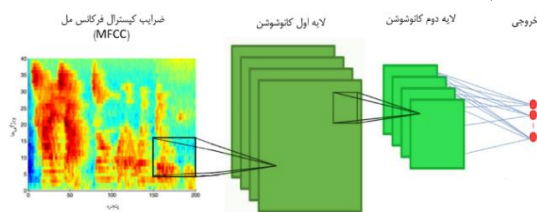


(شکل ۱۴) - طرحی از نظر تحلیلی اشتوتز و همکاران [۳۸]  
(Figure 15) Analytical view for the idea of Stutz et al. [38]

از طرفی آن‌ها بیان کردند که یادگیری ویژگی پدیدآمده توسط چارچوب شبکه تقابلی مولد می‌تواند با مقاومت‌پذیری ارتباط تنگاتنگی داشته باشد. سو<sup>۱</sup> و همکاران در [۶۳] نیز تأکید کردند که بین صحت و مقاومت‌پذیری مصالحه‌ای وجود دارد و بالابردن میزان صحت، کاهش مقاومت‌پذیری را به دنبال دارد. آن‌ها ادعای یادشده را روی ۱۸ معماری محبوب مجموعه دادگان ImageNet پیاده و صحت مصالحه را تأیید کردند. اشتوتز<sup>۲</sup> 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 خصمانه معمولی<sup>۳</sup> و نمونه‌های خصمانه محدود به منیفولد<sup>۴</sup> تقسیم‌بندی و بیان کردند که نمونه‌های خصمانه معمولی روی منیفولد قرار ندارند و مقاومت‌سازی نسبت به 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 خصمانه محدود به منیفولد ناشی از عدم یادگیری صحیح دسته‌بند است و این امر موجب می‌شود تا محدوده تصمیم‌گیری دسته‌بند از محدوده تصمیم‌گیری درست فاصله داشته باشد.

به بیان دیگر، آن‌ها بیان کردند که نمونه‌های خصمانه محدود به منیفولد، ناشی از خطای تعمیم‌پذیری بوده است. تأکید پژوهش آن‌ها بر عدم تداخل بین مقاومت‌پذیری و تعمیم‌پذیری الگوست. آن‌ها تفاوت نمونه‌های خصمانه معمولی از نمونه‌های خصمانه محدود به منیفولد را با محاسبه فاصله بین نمونه اصلی و تقریب تصویرسازی متعامد<sup>۵</sup> نمونه بر منیفولد به دست آوردند.

باتوجه به این که در حمله‌های جعبه سیاه دسترسی به الگو امکان‌پذیر نیست، به‌طور معمول، از الگوی جایگزین برای به دست آوردن اختلال و تولید نمونه خصمانه استفاده می‌شود. شباهت بازنمایی تولیدشده در الگوی اصلی با بازنمایی ویژگی در الگوی جایگزین و همچنین، قابلیت فریب آن‌ها از موضوعات چالشی این حوزه 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] به صورت تقریبی نشان دادند که نمونه‌های خصمانه در زیرفضای همجوار با ابعاد به نسبت بالا قرار دارند. درواقع، آن‌ها بیان کردند که در ابعاد بالاتر بیشتر زیرفضاهای یافت‌شده دارای اشتراک هستند و همدیگر را پوشش می‌دهند. زیرفضای مشترک موجب انتقال‌پذیری نمونه خصمانه از یک الگو به الگوی دیگر می‌شود. آن‌ها بیان کردند که برای وظیفه خاص، معماری‌های مختلف آموزش دیده شبکه عصبی دارای زیرفضاهای مشترک هستند. در دیدگاه آن‌ها، به محدوده تصمیم‌گیری توجه شده است.



(شکل ۱۵) - ساختار یک سامانه پردازش گفتار مبتنی بر استفاده از شبکه‌های عصبی کانولوشنی  
(Figure 16) Structure of a speech processing system using convolutional neural networks

کاریاپا<sup>۷</sup> و قریشی در [۱۳] به مسئله کاهش انتقال‌پذیری نمونه‌های خصمانه پرداختند. آن‌ها تلاش کردند تا با ارائه تابع هزینه تنظیم گرادیان<sup>۸</sup> و استفاده از یادگیری جمعی، موجب کاهش فضای اشتراکی نمونه‌های

<sup>۵</sup> Orthogonal projection

<sup>۶</sup> در پژوهش ترامر و همکاران ذکر شده که این ابعاد حدود ۲۵ بعد

است.

<sup>۷</sup> Kariyappa

<sup>۸</sup> Gradient Alignment Loss (GAL)

<sup>۱</sup> Su

<sup>۲</sup> Stutz

<sup>۳</sup> Regular

<sup>۴</sup> Constrained to the manifold

روش‌هایی نظیر روش شانهر<sup>۴</sup> و همکاران در [۶۶] استفاده شده‌است تا صوت خروجی با عامل انسانی همچنان قابل تشخیص بوده، و تغییرات محسوس نباشد. همین روند در حوزه 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 حوزه پردازش متن آورده شده‌است. مشاهده می‌شود که با بزرگ‌شدن یک حرف، دسته جملۀ موردنظر تغییر کرده‌است. (شکل ۲۳) 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 مجموعه 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 mood 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 opposition Conservatives. 94% Business

(شکل ۱۷) -نمایی از مثال نمونه 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]

## مجموعه دادگان

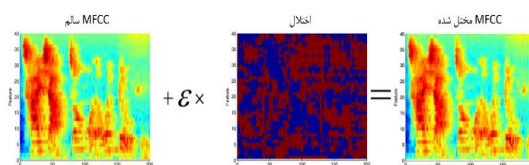
در این پژوهش از مجموعه دادگان MNIST [۶۸] و CIFAR10 [۶۹] استفاده شده‌است. مجموعه داده MNIST متشکل از ۶۰ هزار نمونه اعداد دستنویس انگلیسی و به صورت سطوح خاکستری (تک‌کانال رنگی) است. مجموعه داده CIFAR10 از ۶۰ هزار نمونه تصویر رنگی (سه کانال رنگی) در ده رده مختلف تشکیل شده‌است. تقسیم‌بندی این دادگان‌ها به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

خصمانه بین الگوهای مختلف یادگیری جمعی شوند. همان‌طور که 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 کاریاپا 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 متخصص سخت‌تر الگوی کلی را فریب دهد. از شباهت کسینوسی برای تنظیم کردن گرادین استفاده می‌شود.

## ۵- حمله و دفاع در حوزه‌های غیر از بینایی ماشین

حوزه‌های غیر از بینایی ماشین نظیر پردازش گفتار، متن و... نیز از اختلال‌های خصمانه مصون نیستند. درواقع، افراد متخصص در همه حوزه‌ها به دنبال پیدا کردن اختلال‌های بهینه جهت فریب الگو هستند.

در غالب پژوهش‌های حوزه گفتار با استخراج ضرایب بانک فیلتر ملبرای فریم‌های متوالی، داده صوتی ورودی را به تصویر تبدیل می‌کنند ((شکل ۱۵ و (شکل ۱۶۱۷). در پژوهش ونگ<sup>۱</sup> و همکاران از آن برای تشخیص کلیدواژه صوتی<sup>۲</sup> استفاده کردند. آن‌ها تلاش کردند تا همانند پژوهش‌های حوزه بینایی ماشین به تولید نمونه خصمانه صوتی روی آورده و از این طریق رده کلیدواژه صوتی ورودی را تغییر دهند.



(شکل ۱۶) -نمایی از مختل کردن ضرایب کیسترال

فرکانس مل با استفاده از نمونه‌های خصمانه [۶۵]

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

رویکرد حمله 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<sup>۳</sup> و محاسبه حد آستانه شنوایی نیز در

<sup>1</sup> Wang

<sup>2</sup> Keyword spotting

<sup>3</sup> Force alignment

<sup>4</sup> Schönherr

تقسیم‌بندی انجام‌شده توسط فراهم‌کنندگان هریک از این دادگان‌هاست.

## معماری

در این پژوهش از معماری WRN-28 مطرح‌شده توسط زاگورویکو<sup>۱</sup> و کوموداکیس<sup>۲</sup> استفاده شده‌است [۷۰]. این معماری دارای ۲۸ بلوک کانولوشنی باقیمانده<sup>۳</sup> 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 سیلو<sup>۴</sup> 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، پردازنده گرافیکی GeForce GTX 1080 Ti و حافظه دسترسی تصادفی ۳۲ گیگابایت، همچنین، از زبان پایتون<sup>۵</sup> نسخه ۳/۶ برای رمزنویسی و کتابخانه پایتورچ<sup>۶</sup> [۷۲] نسخه ۱/۶ نیز برای نوشتن بستر آزمایش‌ها استفاده شده‌است.

(جدول ۲) صحت الگوی WRN-28 آموزش‌دیده روی

مجموعه داده MNIST با نمونه‌های خصمانه تولیدشده از

روش‌های حمله مختلف تحت نرم  $l_\infty$

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

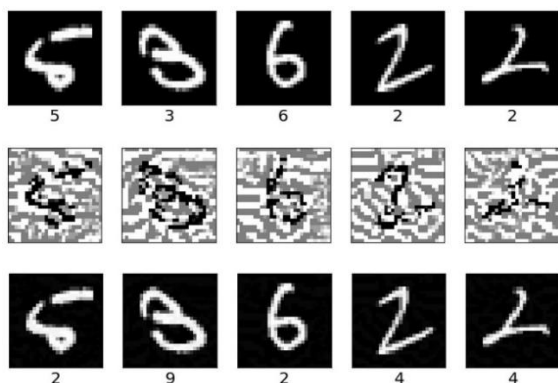
| روش حمله<br>$\epsilon$ | صحت الگوی مینا |       |       |       |       |
|------------------------|----------------|-------|-------|-------|-------|
|                        | ۰/۱            | ۰/۳   | ۰/۱   | ۰/۲   | ۰/۳   |
| FGSM                   | ۹۸/۴۴          | ۹۶/۵۱ | ۷۲/۵۶ | ۲۳/۵۹ | ۱۳/۳۲ |
| PGD                    | ۹۸/۳۹          | ۹۲/۵۲ | ۸/۲۵  | ۰/۲۴  | ۰/۱۰  |
| FFGSM                  | ۹۸/۴۲          | ۹۶/۴۲ | ۸۴/۲۸ | ۳۸/۷۴ | ۱۲/۰۳ |

## مقایسه برخی از پرکاربردترین حملات خصمانه

در ادامه به بررسی چند روش پرکاربرد حمله روی دو معماری یادشده به‌همراه مجموعه دادگان مرتبط با آن‌ها خواهیم پرداخت. 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 نسبت به

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  $l_\infty$  توانسته صحت الگو را کاهش دهد. درواقع، روش PGD به خاطر مرحله‌ای بودن آن و همچنین، تصویرکردن روی محدوده  $l_\infty$  بهتر به مرز دسته‌بند نزدیک شده و به سمت دیگر مرز دسته‌بند هدایت می‌شود. در (شکل ۱۸) نمایی از نمونه‌های سالم، اختلال و نمونه‌های خصمانه 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.

همان‌طور که 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 خصمانه سطر سوم دست پیدا می‌کنیم. همان‌طور که مشخص است، همه برچسب‌های نمونه‌های خصمانه توسط مدل به اشتباه شناسایی شده‌است. چنانچه آزمایش‌های انجام‌شده را روی فضای رنگی مجموعه داده CIFAR10 معماری ذکرشده در بخش ۰ آزمایش کنیم، نتایج صحت حملات مبتنی بر محدوده  $l_\infty$  نظیر FGSM، PGD و FFGSM با میزان اختلال‌های ۰/۰۱ الی ۰/۳ و به الگوی آموزش‌دیده به‌صورت (جدول ۳) خلاصه می‌شود. نتایج حملات مبتنی بر 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

<sup>1</sup> Zagoruyko

<sup>2</sup> Komodakis

<sup>3</sup> Residual

<sup>4</sup> Silva

<sup>5</sup> Python

<sup>6</sup> PyTorch

آن‌ها از طرف دیگر نیز تأثیرگذار بوده و این مدت‌زمان در روش DeepFool به‌خاطر ماهیت سخت‌گیری میزان اختلال به‌مراتب بالاتر است. درواقع، روش DeepFool نسبت به روش‌های دیگر اختلال بسیار کوچکتري را دارد. بحث زمان‌بر بودن محاسبات آن نیز توسط موسوی دزفولی و همکاران در [۱۵] مطرح شده‌است. آن‌ها بیان کردند که رویکردهای دیگر نسبت به DeepFool در مدت‌زمان اجرا حدود ۵ برابر سریع‌تر هستند. در آزمایش‌های انجام‌شده مطابق (جدول ۴) نیز این بیان تأییدشده و روش DeepFool نسبت به حمله CW به‌مراتب در اجرا طولانی‌تر است.

(جدول ۴) - صحت الگوی آموزش‌دیده روی مجموعه

دادگان MNIST و CIFAR10 و زمان اجرای روش‌های حمله

مختلف تحت محدوده 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 |          |
|----------|-------|----------|---------|----------|
| روش حمله | 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 همان‌طور که 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 نسبت به سایر رویکردهای حمله نظیر حمله CW به‌مراتب بیشتر است. البته قدرت حمله الگوریتم DeepFool به‌مراتب از CW بیشتر است. همان‌طور که مشاهده می‌شود، در مقایسه با حمله CW با ۵۰ تکرار، روش DeepFool توانایی کاهش صحت بیشتری نسبت به CW روی مجموعه داده MNIST داشته‌است.

نمونه‌های خصمانه 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

حمله مختلف تحت نرم  $l_\infty$

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

| صحت الگوی مبنا |       |       |       |      |      |
|----------------|-------|-------|-------|------|------|
| روش حمله       | ۰/۱   | 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 |

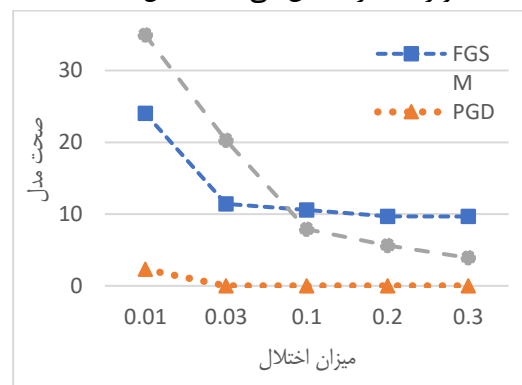
همان‌طور که در

(جدول ۳ آورده‌شده، صحت الگوی آموزش‌دیده روی

مجموعه داده CIFAR10 نسبت به حملات مختلف

محصورشده به نرم  $l_\infty$  نظیر FGSM، PGD و FFGSM با میزان اختلال‌های ۰/۱ 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 چندمرحله‌ای بودن آن است و حتی در میزان اختلال‌های بزرگ‌تر، دوباره روش PGD صحت الگو را بیشتر کاهش می‌دهد (شکل ۲۰).



(شکل ۱۹) مقایسه قدرت حملات محصورشده مبتنی بر نرم  $l_\infty$

روی الگوی آموزش‌دیده با مجموعه داده MNIST

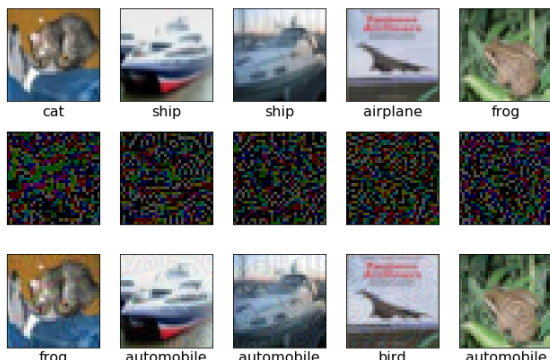
(Figure 20) Comparison of the strength of  $l_\infty$  bounded attacks for the model trained using MNIST dataset

روش‌های مبتنی بر  $l_2$  نظیر CW و همچنین،

DeepFool باتوجه به ماهیت رویکرد حمله آن‌ها به‌طور مجزاً در (جدول ۴ ارزیابی شده‌اند. در این روش‌ها تکرارپذیر بودن آن‌ها از یک طرف و زمان اجرای حمله

<sup>1</sup> Bounded

در (شکل ۲۲) نمایی از نمونه‌های سالم، اختلال تولیدشده توسط روش حمله PGD-50 و همچنین نمونه‌های خصمانه حاصل شده از آن حمله برای الگوی آموزش دیده روی مجموعه داده CIFAR10 آورده شده است.



(شکل ۲۲) - نمایی از نمونه‌های سالم (ردیف بالا)، اختلال (ردیف وسط) و نمونه خصمانه تولیدشده توسط روش حمله PGD-50 (ردیف پایین) برای الگوی آموزش دیده 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

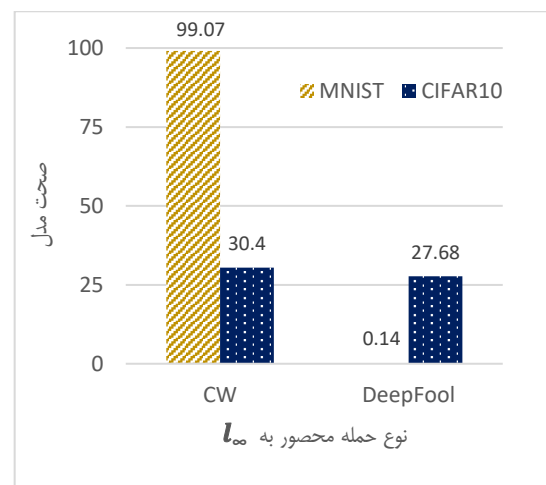
همان طور که در سطر نخست، (شکل ۲۲) مشخص شده، نمونه‌های سالم به درستی توسط الگو شناسایی شده‌اند. با اضافه شدن اختلال حاصل از روش حمله PGD-50، به نمونه‌های خصمانه سطر سوم دست پیدا می‌کنیم. همه برچسب‌های نمونه‌های خصمانه توسط الگو به اشتباه شناسایی شده است. یکی از نکات جالب این است که برای یک رده واحد، با اختلال‌های مختلف PGD-50 روی هر داده تصویری MNIST از برچسب کشتی (ship)، نمونه خصمانه متناظر با آن‌ها به رده خودرو (automobile) تغییر برچسب داده است.

### مقایسه برخی روش‌های کاربرد دفاع

#### نسبت به حملات خصمانه

باتوجه به اینکه رویکرد یادگیری خصمانه به عنوان روشی پایدار و مقاوم نسبت به حملات خصمانه معرفی شده و نسبت به سایر رویکردها برتری داشته، در ادامه طی آزمایش‌هایی تأثیر استفاده از آموزش خصمانه بررسی شده است.

همان طور که در (جدول ۵) مشخص شده صحت الگوی پایه آموزش دیده روی مجموعه داده CIFAR10 نسبت به حملات خصمانه FGSM، PGD و FFGSM با شاخص میزان اختلال ۰/۰۳ نشان داده شده است. زمانی که



(شکل ۲۰) - مقایسه قدرت حملات محصورشده مبتنی بر

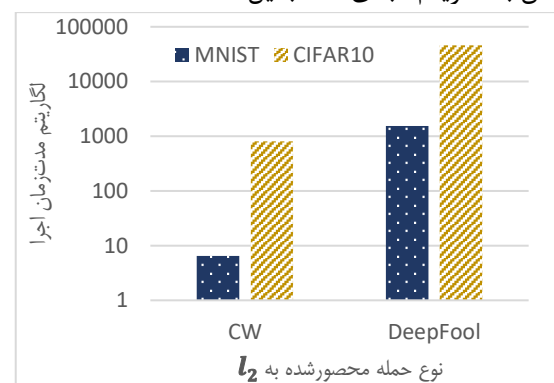
نرم  $l_2$  روی الگوی آموزش دیده با مجموعه داده MNIST و

CIFAR10

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

این نتایج روی مجموعه داده CIFAR10 نیز ارزیابی و نتایج آن در ستون دوم (جدول ۴) درج شده است. همان طور که مشخص شده، مدت زمان اجرای حملات روی مجموعه داده CIFAR10 به واسطه افزایش اندازه از  $28 \times 28$  پیکسل در نمونه داده MNIST به  $32 \times 32$  پیکسل در نمونه داده CIFAR10 طولانی تر شده است. همچنین، مشخص است که مدت زمان اجرای حمله DeepFool نسبت به حمله CW، بسیار طولانی مدت بوده است (در حدود  $46143/5$  ثانیه).

(شکل ۲۱) مقایسه مدت زمان اجرای روش DeepFool و CW روی دو الگوی آموزش دیده روی MNIST و CIFAR10 را نشان می‌دهد. باتوجه به زیادبودن باز مدت زمان اجرای روش DeepFool، میزان زمان به لگاریتم مبنای ۱۰ تبدیل شده است.



(شکل ۲۱) - مقایسه لگاریتم مدت زمان اجرای حملات

محصورشده به  $l_2$

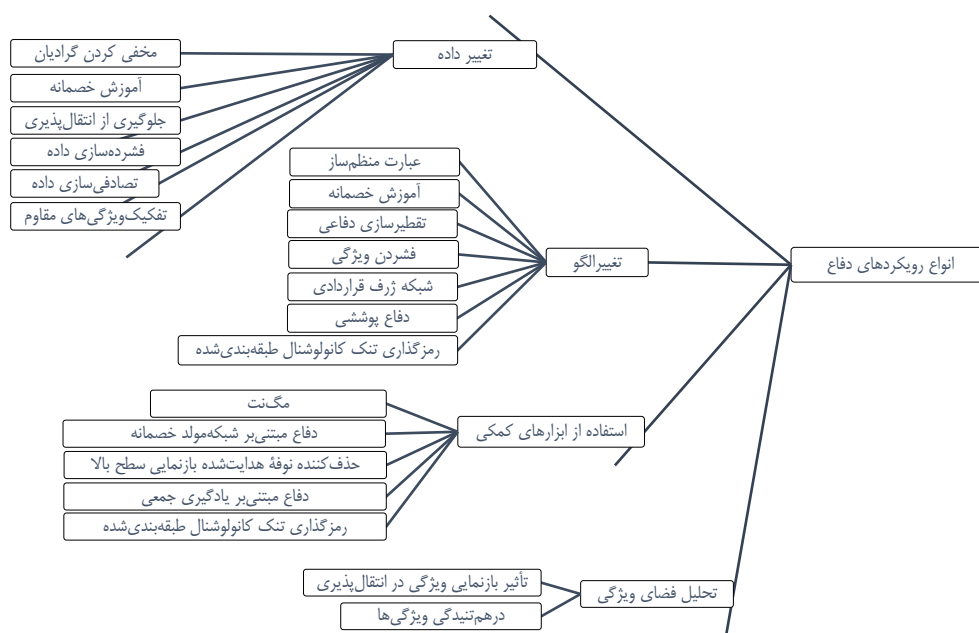
(Figure 22) Comparison of execution time of the  $l_2$  bounded attacks

همین روال آموزش خصمانه با حملات PGD و FFGSM نیز انجام شده و نتایج صحت الگوی آموزش‌دیده روی مجموعه داده آزمون به ترتیب ۹۱/۴۰ و ۷۸/۹ به دست آمده است. مشخص است که صحت در آموزش خصمانه به نسبت کاهش یافته و این امر تا حدودی پدیده منطقی و عادی است. زمانی که به سمت مقاوم‌پذیری حرکت می‌کنیم، لزوماً صحت الگو افزایش ندارد، بلکه کاهش هم خواهد داشت و باید نقطهٔ مصالحه هر دو معیار را شناسایی کرد. الگوی آموزش‌یافته مبتنی بر آموزش خصمانه با حمله FGSM (adv-FGSM)، نسبت به حمله FGSM نسبت به بقیه و به ویژه حالت پایه مقاوم‌تر شده ولی نسبت به سایر حملات همچنان مقاومتی ندارد. از بین همه الگوهای آموزش‌یافته مبتنی بر آموزش خصمانه، الگوی آموزش‌دیده مبتنی بر حمله PGD با تعداد

در حالت عادی شبکه با ۲۰۰ تکرار آموزش می‌بیند، 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 ببیند، صحت الگو روی مجموعه داده آزمون به میزان ۸۳/۵۹ درصد می‌رسد.

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

| 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  |



(شکل ۲۳) خلاصه‌ای از انواع رویکردهای دفاع [۶]، [۱۱] و [۳۸]  
(Figure 24) A summary of different defense approaches

شناسایی کند و روی نقاط حساس هم برای تشخیص برچسب و کلاس متمرکز و از این رو مقاوم‌پذیری آن نسبت به 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، پیدا کردن شاخص بهینه نبوده است. بنابراین، احتمال می‌رود که با شناسایی شاخص‌های بهینه هر روش حمله، هم در مرحلهٔ آموزش

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 نسبت به 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 نسبت به آن‌ها تعریف نماییم، موجب می‌شود که الگوی غالب جوانب مختلفی از بازنمایی داده‌ها را

خصمانه و هم در مرحلهٔ آزمون، شاهد صحت‌های بهتری باشیم. باتوجه به زمان‌بر بودن روش آموزش خصمانه در هر تکرار از آموزش، باید به آن تکه‌داده<sup>۱</sup> حمله شود و چنانچه حمله زمان‌بر باشد، کل آموزش زمان‌بر خواهد شد. این امر در آموزش خصمانه با حمله PGD به مراتب از همه حملات ارزیابی شده در آموزش خصمانه طولانی‌تر بوده است. بنابراین، در روش آموزش خصمانه که به عنوان یکی از برترین رویکردهای دفاع شناخته می‌شود، نوع حملهٔ مورد استفاده 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 نسبت به حملات خصمانه برای تولید الگویی پایدار و قابل اعتماد مهم است. تمرکز و حل چالش‌های این حوزه موجب تولید الگوی مقاوم نسبت به 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 حوزه یادگیری ماشین خصمانه، بخشی از مهمترین مفاهیم آن بازگو و نظرات مختلف به تحلیل و بررسی گذاشته شود. پس از ارائه مفاهیم کلیدی این حوزه به بررسی و نقد رویکردهای متنوع حمله به شبکه‌های عصبی عمیق پرداخته شد. مشخص شد که حملات تکرارشونده از حملات عادی الزاماً بهتر نیستند، ولی تعریف مسئله روش DeepFool در  $l_2$  و همچنین، PGD در  $l_\infty$  از رویکردهای سابق قدرت بیشتری دارند. در بخش دیگری از این مقاله به بیان تعریف دفاع از شبکه‌های عصبی عمیق نسبت به نمونه‌های خصمانه و همچنین، به مرور و نقد برخی از مهمترین رویکردهای دفاع پرداختیم.

<sup>1</sup> Batch data



- 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 (ICSS), 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].

رویکرد آموزش خصمانه را به عنوان 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 فراگرفته شده در شبکه های عصبی عمیق پرداختیم و بر اساس پژوهش های موجود بیان کردیم که مصالحه میان تعمیم پذیری و صحت موثر بوده و الزاماً این دو مفهوم در تلاقی با یکدیگر نیستند. در بخش آخر نیز، به انجام آزمایش هایی جهت بررسی رویکردهای مختلف حمله خصمانه و بررسی قدرت هریک پرداختیم. بیان شد که در ناحیه محصور شده  $l_{\infty}$  روش PGD و در ناحیه محصور شده  $l_2$  روش DeepFool قدرت بیشتری دارند. زمان اجرا از نکاتی بود که تحلیل، و مشخص شد که باتوجه به برتری روش های PGD و DeepFool، مدت زمان بیشتری نسبت به سایر هم ترازهای خود برای اجرا نیاز دارند و این میزان در DeepFool از همه روش های حمله بیشتر بود. همچنین، به مقایسه برخی از رویکردهای پرکاربرد دفاع نسبت به نمونه های خصمانه نیز پرداخته شد و از بین روش های مبتنی بر نواحی حول داده  $l_{\infty}$ ، روش آموزش خصمانه مبتنی بر PGD با شاخص های مربوط از سایر روش ها بهتر بوده و در مقابل اغلب روش های حمله مقاوم عمل کرده است. گفتنی است که روش های مختلف حمله خصمانه و همچنین، روش های مختلف دفاع نسبت به 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](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.

\* Corresponding author

\* نویسنده عهده دار مکاتبات

سال ۱۴۰۲ شماره ۲ پیاپی ۵۶

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

فصلنامه

پژوهش های نوین در حوزه های نوین

۱۳۹

- [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
- [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 Conference on 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.

- 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 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

زبان طبیعی، یادگیری ماشین، یادگیری عمیق و همچنین، طراحی سخت‌افزار. نشانی رایانامه ایشان عبارت است از:

homayoun@aut.ac.ir



مریم امیرمزلقانی در سال ۱۳۸۸ در رشته مهندسی برق از دانشگاه صنعتی 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 موردعلاقه ایشان عبارت است از: پردازش سیگنال آماری، پردازش تصویر، پردازش سیگنال چندرزولوشنی و اخفای اطلاعات. نشانی رایانامه ایشان عبارت است از:

mazlaghani@aut.ac.ir

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 Proceedings 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 خالویی از سال ۱۳۹۶ به



عنوان دانشجوی دکتری مهندسی کامپیوتر در گرایش هوش مصنوعی دانشگاه صنعتی امیرکبیر تحت نظارت

دکتر محمد مهدی همایون‌پور و دکتر مریم امیرمزلقانی مشغول به تحصیل شد. زمینه‌های پژوهشی موردعلاقه ایشان عبارت است از: یادگیری عمیق، یادگیری ماشین خصمانه، شبکه‌های مولد و بهینه‌سازی. نشانی رایانامه 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 موردعلاقه ایشان عبارت است از: پردازش گفتار، پردازش سیگنال، پردازش

فصل بی



