پژوهشکده علوم کامپیوتر

پروپوزال همکاری پژوهشی تک پروژه غیر مقیم

حملات سایبری به سیستمهای هوشمند مولد: تحلیل و ایجاد مقاومت در برابر آنها

Cyber Attacks to Smart Generative Systems: Analysis and Development of Countermeasures

نام و نام خانوادگی متقاضی:	نشانی: تهران، خیابان کارگر شمالی، تقاطع خیابان	نشانی پست الکترونیکی:
دکتر محمد صیادحقیقی	جلال آل احمد، دانشکده فنی دانشگاه تهران، دانشکده برق و کامپیوتر	sayad@ut.ac.ir
شمارهٔ تلفن ثابت: ۸۸۶۸۴۵۷۵	شمارهٔ تلفن همراه: ۰۹۱۲۲۰۲۷۲۵۱	

چکیده

ارایه خدمت اکنون در بسیاری از بخشها با کمک هوش مصنوعی و یا سیستمهای هوشمند انجام میشود. حتی دامنه کاربرد این سیستمهای هوشمند، نوید گذار از Industry 4.0 به Industry 5.0 را به ارمغان آورده است. در مورد یک سیستم هوشمند مجرد (برای مثال، یک سامانه مبتنی بر یک مدل هوش مصنوعی عمیق)، دشمن میتواند عملکرد مدل را هدف بگیرد تا سیستم کار غیر رضایت بخشی را انجام دهد و یا می تواند حریم خصوصی داده های مربوط به آن مدل را (برای مثال با یافتن نمونه های آموزشی) مورد تعرض قرار دهد. الگو های تامین امنیت کلاسیک با روشهای رمزنگارانه، لزوماً در این اکوسیستم جدید کاربرد ندارند، و یا حداقل بخشی از راه حل امنیتی هستند. در این پروژه سعی میشود که راهبرد های حمله و دفاع در سیستمهای هوشمند، بخصوص سیستمهای مولد مبتنی بر شبکه های عصبی عمیق و یادگیری ماشین، مورد بررسی قرار گیرد. در این راستا، هم در بخش تهاجمی به توسعه مدلهای جدید پرداخته خواهد شد و هم در حوزه دفاعی، برای چنین حملاتی روشهای مقابله موثر توسعه داده خواهد شد. نتایج اولیه مطلوبی در این دو حوزه هم اکنون در مورد حمله گریز حاصل شده است.

واژگان کلیدی:

امنیت سایبری، حریم خصوصی، شبکه عصبی عمیق، سیستمهای هوشمند، حمله گریز، یادگیری ماشین متخاصم

Abstract:

New service delivery models now encompass the use of artificial intelligence and intelligent systems in many sectors. The wide deployment of these intelligent systems has even paved the way for the transition from Industry 4.0 to Industry 5.0. In the case of a standalone intelligent system (e.g., a system based on a deep neural network), an adversary can target the model's performance to cause undesirable behavior or compromise the privacy of the data associated with the model (e.g., by training set membership inference). Traditional security models relying on cryptographic methods do not necessarily apply in this new ecosystem, or are at best only part of the security solution. This project aims to examine attack and defense strategies in intelligent systems, particularly generative adversarial neural networks. In this regard, new models will be developed in the offensive domain, while in the defensive domain, effective countermeasures will be designed against such attacks. Initial promising results in both areas have already been obtained for the evasion attack.

Keywords: Cyber Security, Privacy, Deep Neural Network, Smart Systems, Evasion Attack, Membership Inference, Adversarial Machine Learning

مقدمه

مدلهای جدید ارایه خدمات، اکنون در بسیار از بخشها شامل استفاده از هوش مصنوعی و سیستمهای هوشمند میشود. حتی دامنه این کاربرد سیستمهای هوشمند، نوید گذار از Industry 4.0 به Industry 5.0 را به ارمغان آورده است [۱, ۲].
تفاوت این تحول با مدلهای قدیمی تر، هوشمند شدن و قابلیت ارتباط گیری خودکار اجزاء درگیر در سامانه های صنعتی است. این مفاهیم در عمل نزدیک به ویژگیهایی است که در اینترنت اشیا (IoT) و سیستمهای سایبری-فیزیکی (CPS) یافت می شود. اما این تکنولوژی ها علاوه بر مزایای خود، چالش های امنیتی بی سابقه ای را به همراه آورده اند. سامانه های هوشمند، مانند هر فناوری با نفوذ گسترده دیگر، مسائل مربوط به امنیت خود را نیز به همراه خواهد آورد.
در مورد یک سیستم هوشمند مجرد (یک سامانه مبتنی بر یک مدل هوش مصنوعی عمیق)، اولین هدف دشمن میتواند خراب کردن عملکرد مدل باشد تا کار غیر رضایت بخشی را انجام دهد. دومین هدف میتواند نقض حریم خصوصی باشد. برای مثال، در حمله اول، یک مهاجم ممکن است مدلی را که قرار است پرتره‌های انسانی ایجاد کند، به مدلی تبدیل کند که تصاویر نامناسب تولید کند، یا اگر مدل طبقه بند باشد، به جای طبقه‌بندی صحیح تصاویر به عنوان پاندا، ممکن است آنها را به عنوان سگ طبقه‌بندی کند. در خانواده دوم حملات، نقض حریم خصوصی ممکن است شامل سرقت داده های آموزشی یا کل مدل آموزش دیده باشد. یک مثال معروف از آن جایی بود که دشمنان مدل‌های آموزش دیده آمازون را از طریق پرسش‌های جعبه سیاه از طریق API های ارائه‌شده توسط پلتفرم، کپی کردند. از همین روش برای بازیابی مجموعه آموزشی به منظور کسب اطلاعات خصوصی نیز در گذشته استفاده شده است [۳].
الگو های تامین امنیت کلاسیک با روشهای رمزنگارانه، لزوماً در این اکوسیستم جدید کاربرد ندارند، و یا حداقل بخشی از راه حل امنیتی هستند (مانند استفاده از روشهای رمزنگاری همریخت در یادگیری فدرال). حملات به سیستمهای هوشمند، پیچیده و گاهی ترکیبی هستند. به همین جهت لازم است که برای تحلیل و مقابله با حملات و تهدیدات شبکه ها و سیستمهای جدید هوشمند، از روشهای پیچیده تر تامین امنیت که وابسته به زمینه/موضوع هستند استفاده کنیم. در این پروژه سعی میشود که راهبرد های تحلیلی و دفاعی در امنیت سیستمهای هوشمند، بخصوص سیستمهای هوشمند مولد مبتنی بر شبکه های عمیق و یادگیری ماشین، مورد بررسی قرار گیرد. مدل‌های مولد عمیق یکی از انواع مدل‌های یادگیری عمیق هستند که با توجه به توانایی آن‌ها در تولید داده‌ها برای برنامه‌های مختلف مانند مراقبت‌های بهداشتی، فناوری مالی، نظارت و بسیاری موارد دیگر، توجه زیادی را به خود جلب کرده‌اند. محبوب‌ترین این مدل‌ها شبکه‌های متخاصم مولد (GAN) و خودکدگذارهای متغیر (VAE) هستند. در همه مدلهای یادگیری ماشین، نگرانی در مورد نقض امنیت و نشت حریم خصوصی وجود دارد و مدل‌های مولد عمیق نیز از این قاعده مستثنی نیستند. در واقع، این مدل‌ها در سال های اخیر به سرعت پیشرفت کرده‌اند اما کار بر روی امنیت آنها هنوز در مراحل اولیه است. تمرکز ما بر ارتباط درونی بین حملات و معماری‌های مدل و به‌طور خاص، بر روی پنج مولفه مدل‌های مولد عمیق شامل: داده‌های آموزشی، کد پنهان، ژنراتورها/کدگشاها، ممیزها/کدگذاران و داده تولیدشده می‌باشد. در صورت اتخاذ مسیر تحلیل حملات روی مدلهای مولد، تمرکز بر روی حملات زمان استنتاج مانند حمله گریز خواهد بود.
با توجه به تجربه قبلی مجری در زمینه امنیت شبکه و یادگیری ماشین و همچنین شبکه محققان همکار بین المللی وی که نتیجه تحقیقات مشترک آنها در این حوزه در بیش از ۵۰ نشریه معتبر Q1 به چاپ رسیده است، مسیر اجرای این پروژه در صورت تامین اعتبار هموار بنظر می رسد.

پیشینه پژوهش

با توجه به چند وجهی بودن تحقیق در زمینه سیستمهای هوشمند، پوشش تمامی پیشینه های تحقیق ممکن نیست. دراین بخش تنها به پیشینه بخش اصلی یعنی سیستمهای هوشمند مبتنی بر شبکه عصبی و آنهم بخش شبکه های مولد پرداخته میشود. حملات مختلفی مدل‌های یادگیری عمیق خصوصا مدل‌های مولد عمیق را تهدید می‌کند که از آن جمله میتوان به حمله گریز[۱] و یا استنتاج عضویت[۲] اشاره کرد.

حملات گریز

حملات شبکه های عصبی و هوشمند، معمولاً به عنوان حملات خصمانه ای شناخته می‌شوند که هدفشان تغییر ورودی‌ها به گونه‌ای است که شبکه‌های عصبی را فریب دهند، در حالی که برای انسان‌ها نامحسوس باقی بمانند. Szegedy و همکاران [۴] اولین بار وجود نمونه‌های خصمانه را نشان دادند که نشان میداد تغییرات جزئی در ورودی‌های مدل می‌تواند باعث تصمیمات نادرست در خروجی شبکه شود. این مطالعه به اهمیت توسعه روش‌های مقاوم در برابر حملات خصمانه اشاره داشت. Goodfellow و همکارانش [۵] روش FGSM[3] را معرفی کردند که با استفاده از گرادیان یا همان مشتق مدل، نمونه‌های خصمانه را با کارایی بالا تولید می کرد. این روش پایه‌ای برای بسیاری از حملات بعدی شد. سپس، Kurakin و همکاران [۶] روش BIM[4] را معرفی کردند که با اعمال مکرر FGSM، نمونه‌های قوی‌تری ایجاد می‌کرد. Carlini و Wagner [۷] مجموعه‌ای از حملات C&W را توسعه دادند که از روش‌های بهینه‌سازی برای تولید نمونه‌های خصمانه مقاوم در برابر روش‌های دفاعی استفاده می‌کردند. Madry و همکاران [۸] حمله PGD[5] را ارائه کردند که اکنون به عنوان یک معیار استاندارد برای ارزیابی پایداری مدل‌ها در برابر حملات خصمانه شناخته می‌شود.

حملات در شکل جعبه سیاه (بدون اطلاع از جزئیات داخل مدل) نیز مورد بررسی قرار گرفته‌اند. Papernot و همکاران [۶] حمله مدل جایگزین را پیشنهاد کردند که نشان می داد نمونه‌های خصمانه قابلیت انتقال به مدل‌های مختلف را دارند. Brendel و همکاران [۹] حمله مرزی[۶] را معرفی کردند که بدون نیاز به گرادیان، نمونه‌های خصمانه را با تنظیم تدریجی ورودی‌ها ایجاد می کرد. Ilyas و همکاران [۱۰] حمله‌ای مبتنی بر بهینه‌سازی تکاملی طبیعی (NES) ارائه کردند که با تعداد پرسش‌های کم، حملات موثری را اجرا می‌کرد. Moosavi-Dezfooli و همکاران [۱۱] حمله مشهور DeepFool را پیشنهاد دادند، که اگرچه از نوع گریز نبود، ولی تغییرات حداقلی را در ورودی‌ها ایجاد کرده و تصمیمات مدل را تغییر می‌داد. Su و همکاران [۱۲] هم نشان دادند که حتی یک تغییر جزئی در یک پیکسل (One-Pixel Attack) می‌تواند عملکرد مدل را تحت تأثیر قرار دهد.

دسته دیگری از حملات که بیشتر جنبه حریم خصوصی را هدف می گیرند، حملات استنتاج عضویت هستند که با هدف تشخیص این که آیا یک نمونه در مجموعه داده‌های آموزشی مدل بوده است یا خیر طراحی شده اند. Shokri و همکاران [۱۳] اولین حمله استنتاج عضویت را پیشنهاد کردند که از مدل‌های سایه برای تقلید رفتار مدل هدف و پیش‌بینی وضعیت عضویت نمونه‌ها استفاده می کرد. Salem و همکاران [۱۴] روش‌های ساده‌تری را ارائه کردند که نشان می داد این حملات حتی بدون مدل‌های سایه بزرگ نیز مؤثر هستند. Nasr و همکاران [۱۵] حملات جعبه سفید را بررسی کردند که در آن،

[1] Evasion
[2] Membership inference
[3] Fast Gradient Sign Method
[4] Basic Iterative Method
[5] Projected Gradient Descent
[6] Boundary Attack

استفاده از گرادیان‌های مدل باعث افزایش دقت حمله می گردید. Yeom و همکارانش [۱۶] تأثیر بیش‌برازش مدل‌ها را بر آسیب‌پذیری در برابرحملات استنتاج عضویت بررسی کردند و نشان دادند که مدل‌هایی با دقت بالا و تعمیم ضعیف، اطلاعات بیشتری در مورد داده‌های آموزشی فاش می‌کنند.

چندین مطالعه، دفاع در برابر حملات استنتاج عضویت را بررسی کرده‌اند. Jayaraman وEvans [۱۷] استفاده از حریم خصوصی تفاضلی را برای کاهش ریسک این حملات بررسی کردند و نشان دادند که تزریق نویز می‌تواند دقت حمله را کاهش دهد. Carlini و همکاران [۱۸] روش‌های ارزیابی بهتری برای این دسته از حملات ارائه دادند و محدودیت‌های روش‌های پیشین را تحلیل کردند Suri و همکاران [۱۹] استراتژی‌هایی مبتنی بر تنظیم مدل را برای افزایش مقاومت در برابر این حملات پیشنهاد کردند. Rahman و همکاران [۲۰] هم بر استفاده از یادگیری فدرال برای کاهش خطر استنتاج عضویت تمرکز کردند و نشان دادند که توزیع داده‌های غیرمتمرکز می‌تواند احتمال حمله را کاهش دهد.

حملات خصمانه به سیستمهای هوشمند، شامل حملات گریز، استنتاج عضویت، و غیره، تهدیدات جدی برای امنیت و حریم خصوصی مدل‌های هوشمند محسوب می‌شوند. در حالی که روش‌هایی مانند آموزش خصمانه و حریم خصوصی تفاضلی می‌توانند به کاهش این تهدیدات کمک کنند، همچنان زمینه تحقیقاتی زیادی هم در زمینه طراحی حمله و هم در حوزه ایجاد مدل‌های مقاوم در برابر این حملات وجود دارد.

دلایل انجام طرح

دلایل ضرورت اجرای این طرح در ادامه آمده است:

با هوشمند شدن سیستمها و شبکه ها، حملات به خود سیستمهای تصمیم گیرنده مبتنی بر هوش مصنوعی و یادگیری ماشین نیز افزایش یافته است. گستردگی کاربرد سامانه های هوشمند، از خودروهای خودران گرفته تا موتور های جستجو و تولید کننده محتوا، سطوح حمله را به شکل بی سابقه ای افزایش داده است. این موضوع ضرورت شناخت حملات این حوزه و نیز طراحی روشهای مقابله با آنها را دو چندان می کند.

اهداف و محصولات طرح

طرح از نوع پژوهشی بوده و منجر به مقالات علمی و یا ثبت اختراع خواهد شد.. امکانسنجی اولیه (Feasibility Study) تحقیق، انجام شده و پاسخهای اولیه بدست آمده است.

برنامه و نحوه اجرای طرح

در مرحله ۱، حملات اصلی به سامانه های هوشمند، بخصوص حملات خانواده گریز به شبکه های هوش مصنوعی و روشهای دفاعی موجود بررسی میگردد.

در مرحله ۲، یک یا چند حمله نوین یا بهبود یافته برای شبکه های هوش مصنوعی و یا یادگیری ماشین بخصوص بر پایه حملات گریز طراحی و آزمایش خواهد شد. برای این کار سعی بر استفاده از روشهای تحلیلی خواهد شد.

در مرحله ۳، یک یا چند مکانیزم دفاعی متناسب با حملات توسعه یافته طراحی خواهد شد به نحوی که در فرآیند تولید حمله و سپس پدافند، به سطح بالاتری از امنیت در سامانه های مبتنی بر هوش دست یابیم.

برنامه زمانی

مدت انجام این طرح، مطابق روال معمول پژوهشکده خواهد بود و در کمتر از ۱ سال به اتمام خواهد رسید.

شماره مرحله / فعالیت	نام مرحله/ فعالیت	درصد وزنی به کل پروژه	مدت زمان اجرا (ماه)	نمودار زمان‌بندی(ماه)											
				۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۱	۱۲
۱	بررسی روشهای حمله و دفاع فعلی در سامانه های هوشمند	٪۳۰	۲												
۲	طراحی حملات توسعه یافته مبتنی بر روشهای تحلیلی	٪۵۰	۷												
۳	توسعه روشهای دفاعی متناسب با حملات توسعه یافته و تست و شبیه سازی آنها	٪۲۰	۳												

همکاران طرح

همکار علمی طرح، دکتر فائزه فریور عضو هیئت علمی دانشگاه آزاد، واحد علوم و تحقیقات تهران است. سایر همکاران خارجی مانند دکتر مامون العزب نیز ممکن است در جریان تحقیق کمک نمایند.

ابزار مورد نیاز

کیفیت انجام پروژه و میزان واقعی بودن بستر تست و ارزیابی روشهای توسعه داده شده، به مقدار بودجه تصویب شده برای پروژه بستگی دارد. بطور ایده آل این پروژه نیاز به بستر های یادگیری ماشین با پردازنده های گرافیکی قوی نیاز دارد. برای کار در مقیاس کوچک، با توجه به اینکه بسیاری از نرم افزارها قابل حصول بصورت آزمایشی و یا رایگان هستند، می توان در بخش تولید مقاله از آنها استفاده نمود.

بودجه تقریبی

میزان کار لازم برای این طرح در بخش پژوهشی حدود ۱۸ نفر ماه است که هزینه ای در حدود ۱/۰۸۰/۰۰۰/۰۰۰ ریال برای آن پیش‌بینی شده است. اما هر گونه حمایت پژوهشگاه دانش های بنیادی کمک شایانی خواهد بود. هزینه ها به نسبت درصد وزنی در فازها توزیع می‌شوند.

مراجع

[١] C. Dong, M. Shafiq, M. M. Al Dabel, Y. Sun, and Z. Tian, "DNN Inference Acceleration for Smart Devices in Industry 5.0 By Decentralized Deep Reinforcement Learning," *IEEE Transactions on Consumer Electronics*, 2023.

[٢] C. Trivedi *et al.*, "Explainable AI for Industry 5.0: Vision, Architecture, and Potential Directions," *IEEE Open Journal of Industry Applications*, 2024.

[٣] H. Sun, T. Zhu, Z. Zhang, D. Jin, P. Xiong, and W. Zhou, "Adversarial attacks against deep generative models on data: a survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 4, pp. 3367-3388, 2021.

[٤] C. Szegedy, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[٥] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[٦] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*: Chapman and Hall/CRC, 2018, pp. 99-112.

[٧] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 ieee symposium on security and privacy (sp)*, 2017: Ieee, pp. 39-57 .

[٨] A. Mądry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *stat*, vol. 1050, no. 9, 2017.

[٩] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," *arXiv preprint arXiv:1712.04248*, 2017.

[١٠] A. Ilyas ,L. Engstrom, A. Athalye, and J. Lin, "Black-box adversarial attacks with limited queries and information," in *International conference on machine learning*, 2018: PMLR, pp. 2137-2146 .

[١١] 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*, 2016, pp. 2574-2582 .

[١٢] 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.

[١٣] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *IEEE symposium on security and privacy (SP)* ,2017 :IEEE, pp. 3-18 .

[١٤] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, "Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models," *arXiv preprint arXiv:1806.01246*, 2018.

[١٥] M .Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning," in *Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP)*, 2018, vol. 2018, pp. 1-15 .

[١٦] S. Yeom, I. Giacomelli, A. Menaged, M. Fredrikson, and S. Jha" ,Overfitting, robustness, and malicious algorithms: A study of potential causes of privacy risk in machine learning,"

Journal of Computer Security, vol. 28, no. 1, pp. 35-70, 2020.

[۱۷] B. Jayaraman and D. Evans, "Evaluating differentially private machine learning in practice," in *28th USENIX Security Symposium (USENIX Security 19)*, 2019, pp. 1895-1912 .

[۱۸] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer, "Membership inference attacks from first principles," in *IEEE Symposium on Security and Privacy (SP)*, 2022: IEEE, pp. 1897-1914 .

[۱۹] A. Suri, P. Kanani, V. J. Marathe, and D. W. Peterson, "Subject membership inference attacks in federated learning," *arXiv preprint arXiv:2206.03317*, 2022.

[۲۰] M. M. Rahman, A. S. Arshi, M. M. Hasan, S. F. Mishu, H. Shahriar, and F. Wu, "Security risk and attacks in AI: A survey of security and privacy," in *IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, 2023: IEEE, pp. 1834-1839 .

نام و نام خانوادگی متقاضی: دکتر محمد صیادحقیقی– دانشیار دانشگاه تهران

تـــــاریخ: ۱۴۰۳/۱۱/۱۵