Research Article, 2021, Volume 8, Number 1
http://www.jcomsec.org

Journal of Computing and Security

Persian Abstract

روش یادگیری خودنظارتی تقابل گشتاور مبتنی بر آموزش برای حملات خصمانه

منیره مشوش [آ]، دکتر مهدی افتخاری [آ،*]، کاوه بهرامن [آ]

[آ] بخش مهندسی کامپیوتر- دانشگاه شهید باهنر کرمان- کرمان- ایران

اطلاعات مقاله

تاریخچه مقاله:
دریافت: 5 January 2021
اصلاح: 2 June 2021
پذیرش: 29 June 2021
انتشار آنلاین: 21 July 2021

کلمات کلیدی:
یادگیری خصمانه، پیش‌بینی شیب نزولی، دفاع، دقت قوی، یادگیری عمیق

چکیده

با پیشرفت سریع یادگیری عمیق و استفاده از آن در انواع کارها، شبکه‌های عمیق نشان داده‌اند که در برابر نمونه‌های خصمانه آسیب‌پذیر هستند. تحقیقات اخیر نشان داده است که استفاده از یادگیری خود نظارت شده به روش‌های مختلف منجر به افزایش مقاومت شبکه می‌شود. این مقاله به بررسی تأثیر نوع خاصی از یادگیری خود نظارتی متضاد به نام تقابل گشتاور (موکو) در افزایش مقاومت شبکه در برابر نمونه‌های خصمانه می‌پردازد. برای این منظور، موکو به عنوان یک کار پیش متن برای پیش آموزش شبکه استفاده می‌شود. سپس تنظیم دقیق باعث افزایش قدرت شبکه در برابر نمونه‌های خصمانه می‌شود. یک روش حمله‌ی جدید مبتنی بر موکو و یکی از روش‌های پیش‌بینی شیب نزولی یا علامت گرادیان سریع معرفی شده است که به داده‌های برچسب دار نیازی ندارد. با استفاده از این داده‌های خراب و روش آموزش خصمانه، به پیش آموزش یک شبکه‌ی عمیق پرداخته می‌شود. سپس از ارائه‌های بدست آمده برای تنظیم دقیق کارهای پایین دستی که منجر به افزایش مقاومت شبکه می‌شود، استفاده می‌شود. به عنوان مثال، با ساختار Resnet50، حمله‌ی پیش‌بینی شیب نزولی و موکو نسخه‌ی ۱ به ترتیب پیشرفت‌های ۲.۷۹٪ ، ۲٪ و ۱.۳۵٪ در مقایسه با روش‌های جیگسا، چرخش و سلفی حاصل شده است. جزئیات بیشتر آزمایشات و پیشرفتهای به دست آمده توسط موکو در قسمت نتایج ارائه شده است و برتری مدل‌های مبتنی بر موکو را در مجموعه داده‌های CIFAR-10 و CIFAR-10-C نشان می‌دهد. همچنین، نتایج به دست آمده برای اعتبار سنجی مدل‌های پیشنهادی در برابر نویزهای مختلف با قدرت‌های مختلف خرابی، مقاومت روش‌های پیشنهادی را تأیید می‌کند.

* نویسنده مسئول.
آدرس‌های رایانامه: monireh.moshavash@gmail.com (م.مشوش)، m.eftekhari@uk.ac.ir (م. افتخاری)، sadegh.b218@gmail.com (ک. بهرامن)

ISSN: 2322-4460