Akademici z FIT ČVUT vypublikovali článek v časopisu Nature Scientific Data

14.1.2025 - Viktorie Dittrichová | Fakulta informačních technologií ČVUT v Praze

Důležitost modelů strojového učení pro detekci bezpečnostních hrozeb na počítačových sítích je dlouhodobě známá jak vědecké, tak odborné komunitě. Výzkumníci sdružení CESNET zkoumají využití metod strojového učení nad síťovým provozem v rámci projektu "Analýza šifrovaného provozu pomocí síťových toků". Ačkoliv bylo již během projektu vytvořeno několik vysoce inovativních a přesných detektorů strojového učení, jejich masovému nasazení stále brání několik obtížně řešitelných problémů. Jedním z nejčastěji zmiňovaných je problém tzv. datového posunu - jev, kdy model strojového učení byl vyvinut na datech, která zastarala a již neodrážejí aktuální stav.

Datové sady v běžném životě a jejich fungování

Možná jste se někdy setkali se situací, kdy jste se pokoušeli přihlásit do svého telefonu nebo počítače pomocí rozpoznání obličeje (například Apple Face ID nebo Windows Hello), ale zařízení vás jednoduše nepoznávalo. To se stalo, protože systém byl natrénován na vaší historické podobě, která se mohla změnit—například kvůli probdělé noci došlo mírnému otoku obličeje, nebo jste změnili účes, který nyní zasahuje jinak do obličeje. V takovém případě se projevil datový posun; trénovací data (vaše podoba) byla zastaralá a ověření nefungovalo správně.

Biometrické ověření obličeje však efektivně čelí problému datového posunu prostřednictvím pravidelného přetrénovávání. Pokaždé, když zařízení úspěšně ověří váš obličej, aktualizuje vaši podobu, aby vás příště opět rozpoznalo. Tento systém obvykle funguje, protože naše podoba se mění relativně pomalu. Nicméně, pokud dojde k náhlé změně (například pokud se muži oholí), ověření často selhává a je nutné aktivovat záložní metodu—zadání hesla.

Důležitost datových sad pro bezpečnost síťového provozu

Podobný problém se objevuje také v oblasti kybernetické bezpečnosti. Na rozdíl od většiny běžných situací je však datový posun v kybernetické bezpečnosti obvykle náhlý a nepředvídatelný. Kyberzločinci mohou nalézt nové metody útoků, nebo nasazení nových služeb na síti může dramaticky ovlivnit charakter provozu. I drobné aktualizace certifikátů mohou zásadně změnit charakter síťových dat a tím narušit funkčnost strojového učení.

V oblasti kyberbezpečnosti obvykle nemáme záložní metody detekce, které by fungovaly na 100 %, a proto je zásadní tento fenomén zkoumat. Vzhledem k faktické absenci dostupných datových sad vhodných pro tento výzkum měli vědci dosud omezené možnosti—naštěstí právě vznikla nová datová sada, která tento výzkum umožňuje.

https://fit.cvut.cz/cs/zivot-na-fit/aktualne/zpravy/21813-akademici-z-fit-cvut-vypublikovali-clanek-v-casopisu-nature-scientific-data