Niezawodne i Bezpieczne Modele Generatywne

Niniejsza praca przedstawia zestaw rozwiązań wspierających rozwój wiarygodnych i bezpiecznych modeli generatywnych. Proponujemy nowe podejścia do modelowania generatywnego, ukierunkowane na praktyczne zastosowania, ze szczególnym uwzględnieniem fizyki wysokich energii. Jednocześnie wprowadzamy metody chroniące własność intelektualną zawartą w modelach i danych treningowych. Dzięki temu wzmacniamy bezpieczeństwo i niezawodność generatywnego uczenia maszynowego, wspierając zarówno twórców i użytkowników modeli, jak i właścicieli danych.

W pierwszej części pracy koncentrujemy się na tworzeniu wiarygodnych modeli generatywnych na potrzeby zastosowań naukowych. Szczególny nacisk kładziemy na symulację eksperymentów z zakresu fizyki wysokich energii prowadzonych w Europejskiej Organizacji Badań Jądrowych (CERN). Proponujemy rozwiązania oparte na uczeniu maszynowym jako alternatywę dla tradycyjnych metod symulacji stosowanych w Wielkim Zderzaczu Hadronów. Osiągnięte rezultaty obejmują opracowanie generatywnych sieci adwersarialnych wiernie odwzorowujących różnorodność danych treningowych, a także stworzenie modelu opartego na mieszance ekspertów, umożliwiającego uchwycenie wielomodalnej natury generowanych danych.

W drugiej części pracy skupiamy się na ochronie wartości intelektualnej w modelach uczenia głębokiego. Wraz z rosnącą zdolnością modeli do rozwiązywaniu konkretnych problemów wzrasta również ich wartość. Naturalnie tworzy to potrzebę opracowania metod jej ochrony. W odpowiedzi proponujemy pierwszą metodę obrony przed nieautoryzowanym odtwarzaniem modeli typu enkoder, pozwalającą wykrywać próby eksploracji przestrzeni odpowiedzi modelu przez atakującego.

Ostatnia część pracy rozszerza perspektywę ochrony wartości intelektualnej z modeli na dane. Pokazujemy, że istniejące podejścia do identyfikacji danych używanych do trenowania modeli dyfuzyjnych mają istotne ograniczenia i proponujemy efektywną metodę wykrywania wykorzystania zbiorów chronionych prawem autorskim. Jako pierwsi analizujemy również zagrożenia dla prywatności w autoregresyjnych modelach wizyjnych i skutecznie adaptujemy naszą metodę do tego typu architektur.

Podsumowując, niniejsza praca wnosi wkład w rozwój modeli generatywnych zdolnych w sposób niezawodny odpowiadać na rzeczywiste wyzwania naukowe, a jednocześnie dostarcza mechanizmy ochrony zarówno modeli, jak i danych. Tym samym wspiera tworzenie godnego zaufania, wiarygodnego i odpowiedzialnego ekosystemu uczenia maszynowego opartego na metodach generatywnych.

Słowa kluczowe: Modele Generatywne, Sieci Generatywne Kontradyktoryjne, Modele Dyfuzyjne, Modele Autoregresyjne Obrazów, Fizyka Wysokich Energii, Bezpieczeństwo Uczenia Maszynowego