Rafał Scherer
Katedra Sztucznej Inteligencji
Wydział Informatyki i Sztucznej Inteligencji
Politechnika Częstochowska
Wydział Informatyki
Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie

Częstochowa, dn. 2 kwietnia 2026 r.

Recenzja

rozprawy doktorskiej mgr inż. Jana Dubińskiego, pt.: Reliable and Safe Generative Models.

Niniejszą recenzję opracowano zgodnie z uchwałą nr 146/2025 Rady naukowej dyscypliny informatyka techniczna i telekomunikacja PW. Promotorem jest prof. dr hab. inż. Przemysław Rokita.

1. Charakterystyka tematu, celu i tezy badawczej rozprawy

Praca wpisuje się w dynamicznie rozwijający się obszar głębokich modeli generatywnych, łącząc zagadnienia ich wiarygodnego zastosowania w symulacjach fizycznych z problematyką bezpieczeństwa modeli i ochrony danych treningowych. Rozprawa ma charakter monograficzny, oparty na cyklu publikacji i obejmuje zarówno oryginalne propozycje metod generatywnych dla zastosowań w fizyce wysokich energii, jak i nowe rozwiązania z zakresu obrony przed kradzieżą modeli oraz identyfikacji danych użytych w treningu dużych modeli dyfuzyjnych i autoregresyjnych. Znaczenie tej rozprawy wynika z dwóch nakładających się trendów: gwałtownego rozwoju dużych modeli generatywnych oraz ich coraz szerszego użycia w zadaniach od symulacji w fizyce wysokich energii po komercyjne systemy wizyjne. Z jednej strony rośnie więc potrzeba wiarygodnych i wydajnych modeli, które można zastosować np. jako zamiennik kosztownych symulacji; z drugiej, te same modele kumulują ogromną wartość intelektualną i prawną, co rodzi pytania o ich ochronę oraz o prawa właścicieli danych treningowych.

2. Zawartość rozprawy

Recenzowana praca mgr inż. Jana Dubińskiego składa się z dziesięciu rozdziałów, wykazu publikacji, bibliografii oraz załączników. Dokument liczy 250 stron.

Pierwszy rozdział przedstawia ogólny kontekst pracy, w której autor zajmuje się zarówno rozwojem wiarygodnych modeli generatywnych dla zastosowań naukowych, jak i ochroną wartości intelektualnej zawartej w modelach oraz danych treningowych. Na początku pokazuje, że współczesne modele generatywne (GAN, modele dyfuzyjne i autoregresyjne) potrafią bardzo dobrze przybliżać złożone rozkłady danych i są już szeroko stosowane w generowaniu obrazów, wideo, mowy czy muzyki, a także w symulacjach fizycznych, projektowaniu leków, medycynie obliczeniowej i nauce o materiałach. Podkreśla, że ta zdolność generowania realistycznych próbek niesie zarówno potencjał, jak i ograniczenia: wciąż trudno jest budować modele, które wychodzą poza proste benchmarki i naprawdę wiernie odwzorowują złożone rozkłady z rzeczywistych zastosowań, a równocześnie rośnie problem ochrony wartości zakodowanej w samych modelach oraz w danych użytych do treningu.

W dalszej części autor zawęża uwagę do dwóch głównych wątków pracy: budowy niezawodnych modeli generatywnych dla fizyki wysokich energii w CERN oraz ochrony modeli i danych przed nadużyciami. W pierwszym obszarze koncentruje się na symulacji odpowiedzi detektorów w Wielkim Zderzaczu Hadronów i proponuje metody oparte na generatywnych sieciach adwersarialnych, które mają zastępować kosztowne symulacje Monte Carlo. Najpierw przedstawia metodę kontrolowanego zwiększania różnorodności próbek generowanych przez GAN, aby lepiej dopasować się do złożonych, wielomodalnych rozkładów danych z kolizji cząstek. Następnie proponuje generatywny model typu mixture-of-experts, który lepiej opisuje wielomodalność danych z eksperymentów wysokich energii.

Kolejna część wprowadzenia pokazuje, że gdy takie modele zaczynają rozwiązywać realne zadania naukowe, stają się jednocześnie cennym zasobem intelektualnym, bo ucieleśniają duże nakłady pracy, danych i mocy obliczeniowej. Autor argumentuje, że w takiej sytuacji samo projektowanie skutecznych modeli nie wystarcza, gdyż trzeba także zadbać o ich ochronę przed nieautoryzowanym kopiowaniem i wykorzystaniem. Stąd drugi duży wątek rozprawy: aktywna ochrona modeli i danych. W części poświęconej modelom autor zapowiada pierwszą aktywną metodę obrony przed kradzieżą enkoderów, polegającą na wykrywaniu intensywnego sondowania przestrzeni odpowiedzi modelu i blokowaniu prób ekstrakcji. W części dotyczącej danych wskazuje na słabości istniejących technik identyfikacji danych treningowych w dużych modelach dyfuzyjnych i proponuje metodę wiarygodnego wykrywania wykorzystania chronionych prawem autorskim zbiorów, a następnie rozszerza ją na nowe modele autoregresyjne obrazów, analizując ich podatność na wycieki prywatności.

W podrozdziale dotyczącym celów badawczych autor formułuje trzy główne zadania: zaprojektowanie wiarygodnych modeli generatywnych dla konkretnych problemów, takich jak symulacje w fizyce wysokich energii, opracowanie zabezpieczeń chroniących wartość intelektualną modeli przed nieautoryzowaną ekstrakcją oraz ochrona wartości danych treningowych i praw ich właścicieli w erze dużych modeli generatywnych. Następnie krótko opisuje, jak poszczególne części pracy realizują te cele: rozdziały 4-5 odpowiadają za część symulacyjną, rozdział 6 za aktywną obronę modeli, a rozdziały 7-9 za ochronę danych i analizę zagrożeń związanych z prywatnością informacji.

Ostatnia część wprowadzenia szkicuje strukturę i logikę dalszych rozdziałów. Autor wyjaśnia, że najpierw przedstawia tło teoretyczne dotyczące modeli generatywnych, szybkich symulacji w CERN oraz ataków i technik ochrony modeli i danych, a także przegląd literatury. Następnie w pierwszej części omawia zastosowanie modeli generatywnych do szybkich symulacji w ALICE i pokazuje ograniczenia standardowych modeli cGAN, wprowadzając SDI-GAN oraz ExpertSim. W drugiej części przechodzi do obrony enkoderów przed kradzieżą danych i opisuje metodę Bucks for Buckets, która monitoruje pokrycie przestrzeni zanurzeń (embedding) przez zapytania i wprowadza kontrolowane zakłócenia, utrudniając ekstrakcję przy zachowaniu użyteczności dla uczciwych użytkowników. W trzeciej części analizuje ochronę danych treningowych w modelach dyfuzyjnych i autoregresyjnych: krytycznie ocenia skuteczność klasycznych ataków inferencji członkostwa, proponuje ramy oceny bardziej realistycznych ataków, wprowadza metodę CDI do identyfikacji chronionych zbiorów w modelach dyfuzyjnych i pokazuje, że modele autoregresyjne obrazów wykazują silną tendencję do zapamiętywania danych treningowych. Rozdział kończy zapowiedź, że całość pracy zmierza do połączenia wysokiej jakości modeli generatywnych z mechanizmami ochrony modeli i danych, aby budować bardziej godny zaufania ekosystem uczenia maszynowego.

Rozdział 2 wprowadza tło teoretyczne potrzebne do zrozumienia dalszych części pracy: opisuje główne typy współczesnych modeli generatywnych, kontekst szybkich symulacji w CERN oraz podstawowe rodzaje ataków i technik audytu modeli i danych. Najpierw autor omawia autoenkodery i wariacyjne autoenkodery jako metodę uczenia reprezentacji i generowania

danych. Następnie przedstawia ideę GAN, ich zalety i problemy z trenowaniem, w szczególności z mode collapse i brakiem enkodera. Kolejna część dotyczy modeli dyfuzyjnych: krokowego zaszumiania i odszumiania danych, uproszczonej funkcji kosztu opartej na przewidywaniu szumu oraz ich rozszerzenia do Latent Diffusion Models, gdzie proces odbywa się w przestrzeni zakodowanej przez autoenkoder, co przyspiesza uczenie i generowanie. Na końcu tej sekcji autor opisuje modele autoregresyjne obrazów oparte na tokenizacji (np. VQ-VAE) i transformerach, a także nowsze warianty (RAR, VAR, MAR), które zmieniają kolejność generowania lub rezygnują z kwantyzacji, co ma wpływ na własności modeli.

Druga główna część rozdziału dotyczy zastosowań modeli generatywnych do szybkich symulacji w CERN, ze szczególnym naciskiem na eksperyment ALICE i detektor Zero Degree Calorimeter. Autor wyjaśnia, że klasyczne symulacje Monte Carlo są bardzo kosztowne obliczeniowo, podczas gdy modele uczenia maszynowego mogą nauczyć się mapowania z parametrów zderzenia na odpowiedź detektora i generować próbki znacznie szybciej. Opisuje rolę ZDC w wyznaczaniu centralności zderzeń, jego budowę jako siatki włókien światłowodowych dających „obrazy" depozycji energii i podkreśla, że rozkłady odpowiedzi są silnie heterogeniczne i wielomodalne, co stawia wysokie wymagania modelom generatywnym. Ten fragment stanowi bezpośrednią motywację dla późniejszych propozycji SDI-GAN i ExpertSim.

W trzeciej części autor wprowadza pojęcie „ochrony wartości" zawartej w modelach i danych i syntetycznie omawia trzy klasy technik: model stealing, membership inference oraz dataset inference. Model stealing opisuje jako proces trenowania modelu-zastępnika na odpowiedziach modelu ofiary, często z użyciem adaptacyjnych zapytań, co pozwala odtworzyć funkcjonalność bez dostępu do wag czy danych. Membership inference definiuje jako rozstrzyganie, czy pojedynczy przykład był w zbiorze treningowym na podstawie zachowania modelu (np. wartości straty), a dataset inference jako rozszerzenie tej idei na całe zbiory; zamiast jednego punktu testuje się, czy cała kolekcja była użyta do treningu, agregując słabe sygnały w mocniejszy wniosek statystyczny. Autor podkreśla, że te techniki mają charakter dwojaki: są jednocześnie formą ataku i narzędziem audytu, które później wykorzystuje w części poświęconej ochronie danych i praw autorskich.

W rozdziale 3 autor omawia literaturę w trzech obszarach, które odpowiadają trzem częściom pracy. Najpierw omawia dotychczasowe zastosowania modeli generatywnych (głównie GAN) do szybkich symulacji w fizyce wysokich energii, pokazując, że istnieją prace nad zastępowaniem klasycznych symulacji, ale wciąż są problemy z wiernym odwzorowaniem złożonych, wielomodalnych rozkładów i stabilnością trenowania. Następnie przedstawia istniejące metody obrony przed kradzieżą nauczonego modelu (*model stealing*), podkreślając, że skupiają się przeważnie na klasyfikatorach lub modelach generatywnych z inną strukturą i nie są projektowane specjalnie dla enkoderów, co uzasadnia potrzebę nowej aktywnej obrony. W trzeciej części omawiane się prace związane z identyfikacją danych treningowych w modelach generatywnych: klasyczne metody typu membership inference, ich adaptacje do dużych modeli dyfuzyjnych oraz pierwsze próby ataków i audytu na poziomie całych zbiorów, wskazując ograniczenia dotychczasowych podejść i miejsce, w które wpasowuje się zaproponowana później metoda CDI oraz analiza prywatności modeli autoregresyjnych obrazów.

Rozdział 4 przedstawia metodę SDI-GAN, która ma poprawić użyteczność modeli cGAN w symulacji ZDC, poprzez kontrolowane zwiększanie różnorodności próbek tylko tam, gdzie dane rzeczywiście są zróżnicowane. Autor wychodzi od obserwacji, że w klasycznych sieciach cGAN w zadaniach fizycznych często pojawia się zjawisko *mode collapse*: generator ignoruje szum i dla danego warunku zwraca w praktyce jedną „typową" odpowiedź, co jest nieakceptowalne przy symulacji zderzeń cząstek. Omówione są istniejące metody zwiększania

różnorodności (np. MS-GAN, DS-GAN, DivCo) i wskazuje ich słaby punkt: narzucają podobny poziom różnorodności dla wszystkich warunków, podczas gdy w rzeczywistych danych CERN rozrzut wyników bardzo zależy od parametrów zderzenia.

Proponowana metoda dodaje do funkcji kosztu prosty człon regularyzujący, który maksymalizuje stosunek odległości między reprezentacjami dwóch wygenerowanych obrazów do odległości między odpowiadającymi im wektorami ukrytymi, ale skalowany lokalną różnorodnością w danych treningowych dla danego warunku *c*. Odległość między obrazami liczona jest nie w przestrzeni pikselowej, lecz w przestrzeni cech na poziomie warstw dyskryminatora, co ma skoncentrować się na różnicach semantycznych, a nie czysto wizualnym szumie. Całkowita funkcja kosztu to klasyczna strata adversarialna plus ważony człon różnorodności z hiperparametrem λ_{div} wpływającym na siłę regularyzacji.

W części eksperymentalnej metoda jest oceniana najpierw na syntetycznym zbiorze 2D, w którym każda klasa ma dwa warianty („spread = False" z małą wariancją i „spread = True" z dużą), a następnie na danych z GEANT4 dla ZDC (ok. 296 tys. przykładów, obrazy 44×44). Na tym zbiorze, zaproponowany model SDI-GAN lepiej niż bazowe modele cGAN, MS-GAN i DivCo, odtwarza różne poziomy rozrzutu dla poszczególnych warunków, podczas gdy konkurencyjne podejścia mają tendencję do uśredniania zachowania. W zadaniu ZDC autor pokazuje, że SDI-GAN zwiększa różnorodność generowanych odpowiedzi tam, gdzie w danych występuje realna wielomodalność, przy zachowaniu zgodności rozkładów energii i innych metryk z symulacją Monte Carlo. Rozdział kończy się konkluzją, że selektywna regularyzacja różnorodności poprawia przydatność modeli cGAN w zastosowaniach naukowych.

Rozdział 5 wprowadza ExpertSim, model typu mieszanina generatywnych ekspertów dla symulacji odpowiedzi ZDC, który ma rozwiązać problem polegający na tym, że pojedynczy GAN (nawet zaproponowany wcześniej SDI-GAN) nie wystarcza do uchwycenia mocno wielomodalnych rozkładów w danych zderzeń. Autor argumentuje, że odpowiedź detektora zależy od wielu czynników (energia, typ wiązki, geometria zderzenia), a różne obszary przestrzeni wejściowej mają inne parametry fizyczne, więc naturalne jest, aby trenować kilka wyspecjalizowanych generatorów zamiast jednego uniwersalnego.

ExpertSim składa się z kilku ekspertów, modeli GAN (ExpertGAN), z których każdy jest trenowany do modelowania części przestrzeni wejściowej, oraz z sieci „routera", która na podstawie warunków zderzenia wybiera lub miesza ekspertów przy generowaniu próbki. Eksperci mają podobną architekturę do SDI-GAN, ale ich zadanie jest węższe; dzięki temu mogą lepiej odwzorować lokalne rozkłady, natomiast ruter uczy się, który ekspert jest odpowiedni dla danego regionu parametrów. Model jest trenowany tak, aby zarówno eksperci, jak i router minimalizowali łącznie stratę adversarialną i dodatkowe człony zachęcające do specjalizacji (np. ograniczające nakładanie się ekspertów).

W części eksperymentalnej autor pokazuje, że ExpertSim poprawia zgodność rozkładów energii względem cGAN i SDI-GAN, redukuje błędy w ogonach rozkładów i lepiej odtwarza strukturę „plam" energii w obrazach ZDC, przy jednocześnie korzystnym czasie inferencji (szczególnie ważnym z punktu widzenia produkcyjnego użycia w CERN). Analizuje też specjalizację ekspertów: każdy z nich pokrywa inny podzakres energii czy typów zdarzeń, oraz wykonuje badania ablacyjne (liczba ekspertów, różne strategie rutingu). Rozdział kończy się stwierdzeniem, że ExpertSim stanowi jakościowy krok naprzód względem pojedynczego modelu GAN i może zastąpić drogie symulacje Monte Carlo w wybranych częściach łańcucha rekonstrukcji.

Rozdział 6 proponuje aktywną metodę obrony przed kradzieżą enkoderów, nazwaną Bucks for Buckets (B4B), zaprojektowaną tak, aby utrudniać atakującym trenowanie kopii enkodera, przy

minimalnym wpływie na jakość reprezentacji zwracanych uczciwym użytkownikom. Autor przyjmuje scenariusz API (typu OpenAI/Cohere), które udostępnia SSL-owy enkoder wizji (np. SimSiam, DINO) i zwraca wektory zanurzeń; pokazuje, że reprezentacje dla typowych zadań downstream zajmują mały, relatywnie spójny podobszar przestrzeni, podczas gdy skuteczny atakujący musi „pokryć" dużą część przestrzeni zanurzeń, by odtworzyć ogólną funkcjonalność enkodera. To prowadzi do kluczowej intuicji: można odróżnić normalne użycie od ataku, monitorując, jak duży fragment przestrzeni reprezentacji zajmują zapytania danego użytkownika.

Metoda B4B składa się z trzech elementów: (1) modułu szacowania pokrycia przestrzeni embeddingów, (2) funkcji kosztu, która zamienia to pokrycie na „karę", oraz (3) transformacji per-użytkownik, które utrudniają ataki typu Sybil (wiele kont). W proponowanej instancji, pokrycie mierzone jest przez lokalne haszowanie wrażliwe na podobieństwo (LSH): każda reprezentacja wpada do zestawu „buckets", a system zlicza, ile bucketów wypełnił dany użytkownik. Funkcja kosztu jest oparta na użyteczności, tzn. dopóki liczba zajętych bucketów jest mała, szum dodawany do reprezentacji jest znikomy; gdy pokrycie rośnie, obrona zaczyna dodawać coraz silniejszy szum do zanurzeń, pogarszając jakość reprezentacji tylko dla użytkowników „eksplorujących" dużą część przestrzeni. Dodatkowo, na każdą tożsamość użytkownika nakładane są odwracalne transformacje, które zachowują użyteczność, ale powodują, że atakujący nie może łatwo scalać reprezentacji z wielu kont w jednym wspólnym układzie bez dodtkowego kosztu. W części eksperymentalnej autor testuje B4B przeciw skutecznym atakom na enkodery SimSiam i DINO, które wykorzystują kontrastowe uczenie lub MSE i augmentacje do zmniejszenia liczby zapytań.

Rozdział 7 bada, na ile realnie da się przeprowadzić membership inference attacks (MIA) na dużych modelach dyfuzyjnych, ze szczególnym naciskiem na Stable Diffusion v1.4. Autor pokazuje, że część wcześniejszych prac raportuje zbyt optymistyczne wyniki, bo korzysta z nierealistycznych założeń: np. z mocnego fine-tuningu na małych zbiorach (co prowadzi do nadmiernego przeuczenia) albo z nienaturalnego doboru zbioru non-members, który różni się rozkładem od danych treningowych. Proponuje zmodyfikowany przebieg eksperymentu: nie modyfikuje oryginalnego SD-v1.4 i konstruuje nowy zbiór LAION-mi, w którym zbiory „members" i „non-members" są starannie dobrane, duplikowane i „zsynchronizowane" tak, by miały możliwie taki sam rozkład. Następnie testuje różne klasy ataków. Autor stwierdza, że pojedynczo-próbkowe MIA na dużych modelach dyfuzyjnych są w praktyce mało wiarygodne i sensowną drogą jest przejście do obecności całych zbiorów uczących, co proponowane jest w kolejnym rozdziale.

Rozdział 8 proponuje metodę Copyrighted Data Identification (CDI) pozwalającą na sprawdzenie, czy dany zbiór (np. kolekcja zdjęć stockowych, portfolio artysty) był użyty do trenowania dużego modelu dyfuzyjnego, takiego jak Stable Diffusion czy DiT. Autor wychodzi od wyniku z rozdz. 7: membership inference dla pojedynczych obrazów daje zbyt słaby sygnał, by wiarygodnie udowodnić naruszenie praw autorskich. CDI przyjmuje więc jako wejście dwie próbki: zbiór podejrzany P (publicznie dostępne obrazy, co do których właściciel chce sprawdzić użycie w treningu) oraz zbiór kontrolny U z tego samego rozkładu, ale z obrazami niepublikowanymi/nieużytymi, a następnie agreguje wielowymiarowe cechy przynależności dla obu zbiorów i porównuje rozkłady. Architektura składa się z trzech etapów: (1) ekstrakcja wielu sygnałów na poziomie próbek; zarówno z istniejących MIAs (denoising loss, SecMI, PIA/PIAN), jak i nowych ręcznie projektowanych cech; (2) nauka oceny modelu, który łączy te cechy w pojedynczy wynik dla każdej próbki; (3) zastosowanie testu t Studenta (lub pokrewnego testu) między rozkładami ocen P i U, co pozwala z wysoką ufnością orzec, czy P ma „podpis" zbioru treningowego danego DM. W eksperymentach na szeregu architektur (LDM, DiT, U-ViT), różnych rozdzielczościach i trybach warunkowania (bezwarunkowe,

klasowe, tekstowe) autor pokazuje, że CDI potrafi przy próbkach rzędu 70 obrazów osiągnąć ponad 99% ufności przy wykrywaniu wykorzystania zbioru w treningu, pozostając odpornym na fałszywe alarmy i częściowe pokrycie (tylko część P faktycznie w treningu).

Rozdział 9 analizuje prywatność wizyjnych modeli autoregresyjnych: VAR, RAR, MAR i pokrewnych w porównaniu z modelami dyfuzyjnymi (DMs), pokazując, że przy podobnej jakości generacji modele IAR ujawniają znacznie więcej informacji o danych treningowych. Autor proponuje nową metodę membership inference attack dla modeli IAR, łącząc idee z ataków na DMs i LLM-y: wykorzystuje tokenową naturę przewidywań (jak w LLM), a także różnicę między inferencją warunkową i bezwarunkową (jak w DMs). Następnie wykorzystuje MIA jako rdzeń dataset inference (DI) zoptymalizowanego pod IAR: dzięki temu, że wszystkie rozważane MIA dobrze działają, nie jest potrzebny etap wyboru najlepszego MIA dla danego zbioru, a wiarygodne DI można przeprowadzić już na około 6 próbkach – dużo mniej niż ~200 próbek wymagane wcześniej dla DI w DMs. Trzecim filarem jest analiza zapamiętywania (*memorization*).

Rozdział 10 podsumowuje cały wkład rozprawy, podkreślając, że jej głównym celem było połączenie wiarygodnych i bezpiecznych modeli generatywnych: najpierw zaproponowano metody, które umożliwiają wierne i szybkie symulacje w HEP (SDI-GAN i ExpertSim), a następnie zestaw technik chroniących zarówno modele (B4B), jak i dane treningowe (realistyczne MIA dla DMs, CDI, ataki na IAR-y). Autor zbiera to w listę szczegółowych udziałów autora i publikacji w doktoracie oraz pokazuje, że praca przesuwa stan wiedzy od generatywnych modeli, po prostu generujących dane, do bardziej dojrzałego ekosystemu, w którym jakość, wydajność i własność intelektualna mają dużą istotność.

W dalszej części szkicuje przyszły rozwój generatywnej SI: przewiduje dalsze zacieranie granicy między modelami dla zastosowań naukowych i komercyjnych, wzrost skali modeli oraz rosnącą wagę pytań o audyt, odpowiedzialność i regulacje, zwłaszcza w kontekście ochrony danych i praw autorskich.

Dalej następuje spis publikacji oraz załączniki.

3. Ocena rozprawy

W ramach rozprawy doktorskiej Pan mgr inż. Jan Dubiński zaproponował zestaw metod, które w sposób oryginalny łączą projektowanie wydajnych i wiarygodnych modeli generatywnych dla fizyki wysokich energii z nowatorskimi metodami ochrony modeli i danych treningowych, tworząc spójny warsztat dla bezpiecznego wykorzystania dużych modeli generatywnych. Tematyka pracy jest bardzo aktualna i potrzebna, oryginalny dorobek doktoranta polega na:

- stworzeniu nowych metody generatywnych dla fizyki wysokich energii: propozycja SDI-GAN, modyfikacji cGAN selektywnie zwiększającej różnorodność próbek, tak by lepiej odwzorować rozkłady danych z detektorów przy zachowaniu jakości oraz opracowanie ExpertSim, mieszanki modeli generatywnych sterowanej ruterem, pozwalającej na szybkie i wierne symulacje w eksperymencie ALICE.
- opracowaniu metod aktywnej obrony przed kradzieżą enkoderów: zaproponowanie Bucks for Buckets, pierwszej aktywnej metody obrony enkoderów, opartej na monitorowaniu pokrycia przestrzeni zanurzeń, adaptacyjnej funkcji kosztu oraz transformacjach przeciw atakom Sybil.
- Realistyczna ocena membership inference na dużych modelach dyfuzyjnych: krytyczna analiza istniejących MIAs dla Stable Diffusion, pokazująca, że wcześniejsze wyniki są często przeszacowane z powodu nierealistycznych założeń oraz zbudowanie zbioru

LAION-mi i ramy eksperymentalnej, które pokazują, że single-sample MIA na dużych DMs są w praktyce słabe i nie wystarczą jako samodzielne narzędzie audytu.

- Zaprojektowanie metody Copyrighted Data Identification (CDI), która łączy wiele sygnałów przynależności na poziomie próbek ze statystycznym testowaniem na poziomie zbioru oraz empiryczne wykazanie, że metoda ta pozwala z wysoką ufnością (>99%) stwierdzić wykorzystanie danego zbioru w treningu dużego modelu dyfuzyjnego już przy stosunkowo niewielkiej liczbie próbek.
- Analiza prywatności wizyjnych modeli autoregresyjnych (IARs) Opracowanie silnych ataków membership i dataset inference dla IAR, pokazujących, że przy podobnej jakości generacji przeciek prywatności jest u nich znacznie większy niż w dyfuzyjnych oraz empiryczne wykazanie szerokiego i niemal dosłownego zapamiętywania oraz omówienie konsekwencji tego faktu dla projektowania i regulacji przyszłych generatywnych modeli wizyjnych.

Rozprawa doktorska uwidacznia wysoką ogólną wiedzę teoretyczną i praktyczną oraz umiejętność prowadzenia pracy naukowej mgr inż. Jana Dubińskiego. Opracował wprowadzenie do tematyki i przegląd literatury związanej z tematyką pracy. Rozprawa doktorska stanowi oryginalne rozwiązanie problemu naukowego. Zaproponowane metody mają duże znaczenie dla nauk technicznych oraz przemysłu, zarówno teoretyczne, jak i aplikacyjne.

Niezależnie od mojej bardzo dobrej oceny pracy, nasunęły mi się następujące pytania i uwagi:

- Metody SDI-GAN i ExpertSim są pokazane głównie na jednym detektorze (ZDC w ALICE) i ograniczonym zestawie warunków eksperymentalnych. Na ile wyniki przenoszą się na inne detektory/geometrie czy energie zderzeń.
- B4B wymaga doboru kilku progów i parametrów (liczba bucketów, kształt funkcji kosztu, parametry transformacji per-user), co może być trudne w rzeczywistym systemie API. Na ile wyniki są stabilne przy innym doborze hiperparametrów.
- Analiza zakłada konkretny styl ataku (globalne pokrycie przestrzeni embeddingów). Można zapytać, jak B4B zachowa się wobec atakującego, który zna mechanizm obrony i próbuje np. kraść model „po kawałku” (lokalny zakres embeddingów) albo tak projektuje zapytania, by długo nie przekraczać progu kosztu.
- Czy może być sytuacja, że w praktyce właściciel danych może mieć trudność z wygenerowaniem naprawdę reprezentatywnego zbioru U w metodzie CDI.

4. Wnioski końcowe recenzji

Podsumowując recenzję stwierdzam, że rozprawa doktorska Jana Dubińskiego, pt.: Reliable and Safe Generative Models” prezentuje oryginalne rezultaty stanowiące rozwiązanie problemu naukowego oraz wkład w rozwój dyscypliny informatyka techniczna i telekomunikacja. Pan Jan Dubiński wykazał się umiejętnością samodzielnej pracy badawczej, znajomością literatury światowej i wiedzą w zakresie uczenia maszynowego, modeli generatywnych i wizji komputerowej. Rozprawa doktorska prezentuje ogólną wiedzę teoretyczną kandydata. Recenzowana praca spełnia wymagania Ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (Dz.U. z 2022 r. poz. 574 z późn. zm.). w dyscyplinie naukowej informatyka techniczna i telekomunikacja. Wnoszę o jej przyjęcie i dopuszczenie do dalszych etapów postępowania doktorskiego. Ponadto ze względu na ponadprzeciętny poziom rozprawy oraz fakt opublikowania prac związanych bezpośrednio z tematyką rozprawy w materiałach najlepszych konferencji klasy A oraz A* w tym obszarze, wnioskuję o wyróżnienie pracy.

Robert Schaefer