Politechnika Łódzka
Instytut Informatyki

Łódź, 27 lutego 2026 roku

prof. dr hab. inż. Adam Wojciechowski
Instytut Informatyki
Wydział Fizyki Technicznej, Informatyki i Matematyki Stosowanej
Politechnika Łódzka
Al. Politechniki 8, 93-590 Łódź

RECENZJA ROZPRAWY DOKTORSKIEJ

Tytuł rozprawy: **Reliable and safe generative models**

Autor rozprawy: **mgr inż. Jan Dubiński**

Promotor rozprawy: **prof. dr hab. inż. Przemysław Rokita**

Jakie zagadnienie naukowe jest rozpatrywane w pracy (teza rozprawy) i czy zostało ono dostatecznie jasno sformułowane przez Autora? Czy tematyka rozprawy jest aktualna lub dostatecznie ważna? Jaki charakter ma rozprawa (teoretyczny, doświadczalny, inny)?

Pan mgr inż. Jan Dubiński podjął w swojej pracy doktorskiej, zagadnienie bezpieczeństwa, niezawodności i użyteczności modeli generatywnych. Dysertacja została zredagowana na podstawie sześciu znakomitych publikacji, przyjętych i wygłoszonych na sześciu różnych wybitnych międzynarodowych konferencjach naukowych: NeurIPS, ICML, CVPR, ECAI, WACV, ICONIP. W szczególności Doktorant brał udział w opracowaniu modeli generatywnych na potrzeby symulacji fizycznych wysokich energii, ochrony modeli uczenia maszynowego przed skopiowaniem oraz identyfikacją danych treningowych wykorzystywanych do uczenia generatywnych modeli dyfuzyjnych i autoregresyjnych.

Tematyka rozprawy jest bardzo aktualna i ważna w kontekście współczesnego rozwoju oraz wykorzystania modeli sztucznej inteligencji. Należy podkreślić, że prace przedstawione przez Doktoranta ewidentnie wyznaczają granice współczesnych osiągnieć naukowych w przedmiotowym zakresie rozprawy.

Rozprawa ma charakter eksperymentalny i nosi w sobie istotny potencjał praktyczny. Doktorant w ramach każdego z poruszanych tematów nie tylko proponował oryginalne i wartościowe rozwiązania, ale popierał je bardzo rzetelną analizą porównawczą i szeroko zakrojoną optymalizacją hiperparametrów, uzasadniając merytorycznie podjęte decyzje. Praca z racji podjętej problematyki, zakresu oraz metodologii badawczej jest pełnoprawnym osiągnięciem o charakterze naukowo-badawczym w dziedzinie nauk inżynieryjno-technicznych, w dyscyplinie Informatyka Techniczna i Telekomunikacja.

Czy w rozprawie przeprowadzono w sposób właściwy analizę źródeł (w tym literatury światowej i stanu zagadnień w przemyśle) świadczącą o dostatecznej wiedzy Autora? Czy wnioski z przeglądu źródeł sformułowano w sposób jasny i przekonywujący?

W przedstawionej do oceny rozprawie doktorskiej, przegląd literatury realizowany był niezależnie w każdym z podjętych wątków, gdyż każda z przytoczonych prac, z racji chronologii badań, aktualizowała niezbędny zakres rozwiązań referencyjnych, odzwierciedlający aktualny stan zagadnienia. Za każdym razem Doktorant odnosił się do najnowszych doniesień literaturowych, wskazywał nisze badawcze lub kontestował skutecznie dotychczasowe osiągnięcia innych naukowców,

zarysowując przydatność i wagę swoich badań. Za każdym razem wnioski formułowano w sposób jasny i przekonujący. Sposób prowadzonego przeglądu świadczy dobitnie o doskonałej wiedzy Autora w poruszanej problematyce.

Przegląd aktualnego stanu zagadnienia niejednokrotnie motywował dobór metodologii badawczej dla prowadzonych eksperymentów. W niektórych pracach doprecyzowywano metodologię, wykazując niedoskonałości poprzednich sposobów weryfikacji. Analiza aktualnego stanu zagadnienia dotyczyła również zbiorów danych wykorzystywanych w eksperymentach. Niejednokrotnie Doktorant konstruował własne zbiory danych, gdyż uznawał, że istniejące zbiory i sposób ich wykorzystania są niewystarczające do prowadzenia rzetelnych eksperymentów. Wszystkie te działania potwierdzają świadomą i dogłębną analizę literatury, a podjęte, na podstawie analizy, działania świadczą o dojrzałości naukowej i stanowią bardzo istotne osiągniecie Doktoranta.

Jak Autor rozwiązał postawione zagadnienie, czy użył właściwej do tego metody?

Pierwsza część dokonań, zaraportowanych w ramach rozprawy doktorskiej, odnosi się do nowych metod wykorzystania modeli generatywnych w wizualizacji zjawisk fizycznych wysokich energii. Dwie pierwsze prace powstały w ramach badań zespołu współpracującego z ośrodkiem CERN, który zajmuje się wizualizacjami trajektorii cząstek elementarnych, w ramach eksperymentu ALICE.

W tym kontekście praca zatytułowana *„Selectively increasing the diversity of GAN-generated samples"*, opublikowana na konferencji ICONIP, porusza problem zwiększenia różnorodności wyników modelu generatywnego, którego celem jest wygenerowanie wiarygodnej wizualizacji trajektorii cząstek elementarnych w odniesieniu do trajektorii rejestrowanych przez detektory. Tradycyjne podejście, bazujące na zmiennych warunkowych, generuje ograniczoną różnorodność trajektorii, co limituje użyteczność tak powstałych symulacji. Opracowano zatem metodę *Selective Diversity GAN* (SDI-GAN), która selektywnie regularyzuje generator w sposób stymulujący różnorodność na etapie trenowania modelu danymi. Istotą metody jest wykorzystanie, do trenowania warunkowego modelu generatywnego *cGAN*, odpowiedniej różnorodności danych treningowych oraz monitorowania dedykowanej funkcji straty, która jest regularyzowana poprzez autorski komponent uwzględniający różnorodność (wariancję) próbek uczących. Dodatkowo, poziom niedopasowania wyników jest mierzony w przestrzeni ukrytej, a nie w przestrzeni pikseli obrazu. Eksperymenty prowadzone były zarówno na danych syntetycznych jak i danych pochodzących z kalorymetru zerostopniowego (*Zero Degree Calorimeter*), używanego w eksperymencie ALICE. Badania wykazały realne przełożenie wariancji danych wejściowych, zakodowanych w funkcji straty, na różnorodność wygenerowanych wyników modelem generatywnym (*cGAN*), bez ograniczania wiarygodności i przy zachowaniu atrakcyjnej wydajności procesu. Porównanie z metodami referencyjnymi wykazało zdecydowaną poprawę wyników symulacji względem wyników literaturowych. Doktorant, jako pierwszy i główny autor, zadeklarował swój wkład w postaci opracowania idei rozwiązania, poprzez implementację i przeprowadzenie eksperymentów, a na analizie wyników kończąc.

Kontynuacja badań w przedmiotowym zakresie zaowocowała powstaniem metody *ExpertSim*, opisanej w artykule pt.: *„ExpertSim: Fast Particle Detector Simulation Using Mixture-of-Generative-Experts"*, przedstawionej na renomowanej konferencji ECAI. Praca kontestuje niedoskonałości poprzedniego rozwiązania (*SDI-GAN*) i zwraca uwagę na ograniczenia pojedynczego generatora w symulowaniu złożonych zachowań cząstek fizyki dużych energii. Zaproponowane rozwiązanie bazuje na modularnej architekturze mieszaniny ekspertów (*mixture of experts*), w której każdy ekspert może koncentrować się na innych dystrybucjach odpowiedzi cząstek – złożone zależności pomiędzy własnościami cząstek i wynikami zarejestrowanymi przez detektory. Architektura w pierwszej kolejności wykorzystuje *router* (wielowarstwowa sieć neuronowa z autorską funkcją kosztu) do przekierowania cech cząstki na jeden

z trzech, wyspecjalizowanych modeli generatywnych (*DCGAN – Deep Convolutional Generative Adversarial Network*) pełniących rolę eksperta. Każdy z ekspertów specjalizuje się w modelowaniu wybranych własności cząstek. Autorzy rozbudowali każdego eksperta o odpowiednią funkcję kosztu względem poprzedniej metody *SDI-GAN*, regularyzując różnorodność, intensywność i lokalizację cząstek na obrazie z detektora. Metodę przetestowano na rzeczywistym zbiorze danych, stosując wiarygodną metodologię. Eksperymenty uzupełniono dogłębną analizą hiperparametrów metody. Uzyskane wyniki jednoznacznie wykazują istotną przewagę rozwiązania nad metodami literaturowymi, ale demonstrują również ogromny potencjał skalowania metody na inne eksperymenty fizyki dużych energii. Doktorant, jako drugi autor artykułu, zadeklarował udział w opracowaniu architektury modelu, stworzenie modeli ekspertów, analizę wyników i udoskonalanie metody.

Drugi wątek prac badawczych, opisany w czterech publikacjach, skupiony jest na zabezpieczeniu wybranych modeli uczenia maszynowego przed skopiowaniem oraz próbą odpowiedzi na pytanie, czy zadany zbór danych, lub wręcz konkretny plik obrazu, był wykorzystany do trenowania wybranych modeli generatywnych, co jest związane z zabezpieczeniem własności intelektualnej zbiorów danych, stanowiących współcześnie podstawę uczenia maszynowego.

W pracy zatytułowanej *„Bucks of Buckets (B4B): Active Defenses Against Stealing Encoders"*, opublikowanej na renomowanej konferencji NeurIPS, opisany jest bardzo interesujący i nowatorski sposób aktywnego zapobiegania eksploracji przestrzeni ukrytej enkoderów. Praca wychodzi z założenia, że zwyczajowe wykorzystanie enkoderów odnosi się do wycinka przestrzeni ukrytej, podczas gdy próba eksploracji szerszego zakresu przestrzeni ukrytej jest próbą niepożądanego skopiowania modelu. Szacowanie poziomu pokrycia przestrzeni ukrytej odbywa się za pomocą zaproponowanej funkcji haszującej (ang. *Local Sensitive Hashing*), która haszuje podobne dane z przestrzeni metrycznej w podobne obszary przestrzeni ukrytej (ang. *hash buckets*). W sytuacji podejrzanego użycia modelu, w sposób adaptacyjny wzrasta wartość zaprojektowanej funkcji kosztu, penalizując próby ekstrakcji danych, poprzez wprowadzenie do odpowiedzi szumu Gaussowskiego o różnych charakterystykach, zależnych od wielkości eksploracji przestrzeni ukrytej. Odporność metody na ekstrakcję przestrzeni ukrytej z wielu kont użytkowników uzyskano poprzez losowe transformacje przestrzeni reprezentacji dla każdego użytkownika (m.in. transformacjami: *Affine, Pad, Shuffle, Add, Binary*). Metoda została przetestowana dla dwóch modeli enkoderów (*SimSiam, DINO*) na popularnych zbiorach danych. Niezwykle wartościowa jest optymalizacja hiperparametrów i dyskusja nad konfiguracją metody. Szeroko zdefiniowane testy wykazały uniwersalność funkcji haszującej dla różnych zbiorów danych. Weryfikacja wykazała zarówno odporność metody na skopiowanie modelu przy dużej liczbie zapytań i wiarygodność dla zwykłych użytkowników adresujących rozsądne liczby zapytań. Testy wykazały również uniwersalność metody względem innych zbiorów danych niż te, na których model był uczony. Próba skopiowania modelu z wielu kont również wykazała skuteczność metody *B4B*, obserwowaną poprzez zmniejszenie jakości odpowiedzi. W przeciwieństwie do rozwiązań literaturowych, *B4B* cechuje się wysoką odpornością na kradzieże modelu przy zachowaniu odpowiedniej jakości dla zwyczajnych użytkowników. Doktorant zadeklarował opracowanie mechanizmu monitorowania eksploracji przestrzeni ukrytej, stworzenie pełnego procesu ataku i obrony, implementację eksperymentów oraz analizę wyników.

Kolejna praca, zatytułowana *„Towards more realistic membership inference attacks on large diffusion models"*, opublikowana na konferencji WACV, porusza kwestie wykrywania nieautoryzowanego użycia danych w procesie uczenia modeli generatywnych. Wiąże się to bezsprzecznie z zabezpieczeniem własności intelektualnej, którą często stanowią same dane. Pytanie, na które próbują odpowiedzieć autorzy to, czy konkretny plik został użyty do trenowania modelu dyfuzyjnego (*stable diffusion*). Praca koncentruje się na stworzeniu dedykowanego zbioru danych *LAION-mi*, który w sposób rzetelny

oddziela podzbiory uczące (*members*) i testowe (*non-members*), zapewniając odpowiednią separację zbiorów i adekwatność rozkładów (*deduplication*, *sanitization*). Autorzy przeprowadzają weryfikację skuteczności wykrywania faktu wykorzystania danych w procesie nauki (MIA - ang. *Membership Inference Attack*) metod literaturowych, wykazując szereg niedoskonałości, w tym brak rzetelności i wiarygodności wcześniejszych opracowań. Stosowana metodologia, w tym dobrane miary jakości, są przekonywujące i ciekawe. Eksperymenty są podstawą wykazania, że istniejące audyty wykorzystania danych są wciąż niewystarczająco zbadane i mało wiarygodne. Autorzy wykazali również ciekawy wpływ douczania modelu (*fine-tuning*) zbiorem POKEMON, który istotnie zaburzał wiarygodność wniosków.

Wnioski z pracy badawczej stworzyły odpowiednią kanwę dla dalszych prac badawczych, które w kolejnym artykule, zatytułowanym *„CDI: Copyrighted Data Identification in diffusion models"*, opublikowanym na renomowanej konferencji CVPR, skoncentrowały się na weryfikacji całych zbiorów danych, a nie pojedynczych próbek. Praca bazuje na założeniu, że słabe przesłanki o wykorzystaniu pojedynczych elementów zbioru mogą być zagregowane w mocną i wiarygodną metodę oceny, czy analizowany zbiór był stosowany do nauki wybranego modelu dyfuzyjnego. Metoda bazuje na autorskiej inżynierii cech. Autorzy uzupełniają literaturowe cechy, tj. *Denoising Loss*, *Step-wise Error Comparing Membership Inference score* (*SecMI*), *Proximal Initialization Attack score* (*PIA*, *PIAN*), o cechy analizujące kluczowy obszar ukrytej reprezentacji modelu, w tym maskowanie 20% gradientów o największym wpływie na stratę. Nowe cechy odzwierciedlają rekonstrukcję straty najbardziej istotnych semantycznie regionów przestrzeni ukrytej i intuicyjnie powinny być niewielkie dla danych użytych w uczeniu modelu (*members*). Dla wzmocnienia sygnału/straty predykcji modelu obliczane jest dziesięć kroków dyfuzji. Autorzy przeprowadzają również optymalizację perturbacji zaszumionej reprezentacji ukrytej modelu, która traktowana jest jako kolejna cecha. Dla zaproponowanego zestawu cech obliczana jest regresja logistyczna dla danych z analizowanych podzbiorów. Odpowiednio skonstruowana metodologia sięga do pełnej reprezentacji ukrytej modelu (*white-box*) lub niepełnej reprezentacji (*grey-box*) oraz bazuje na ujawnionej i ukrytej części zbioru, weryfikowanego względem wykorzystania w nauce modelu. Odpowiednio przeprowadzana analiza statystyczna wskaźników regresji (*membership confidence score*) pozwala z dużym prawdopodobieństwem (*Welch t-test*, $p<0.01$) stwierdzić, że próbki ze zbioru testowego były nielegalnie wykorzystane w trenowaniu modelu dyfuzyjnego. Skuteczność i własności metody *CDI* zostały udokumentowane bardzo bogatymi testami i uzupełniającymi analizami, przeprowadzonymi dla ośmiu popularnych modeli dyfuzyjnych. Skoncentrowano się m.in. na minimalnej wielkości zbioru testowego, który warunkuje poprawne wykrycie, różnych konfiguracjach wektora cech, wpływie udziału danych testowych nie wykorzystywanych do nauki na skuteczność weryfikacji, czy odporności metody na weryfikacje fałszywie pozytywne. Doktorant, jako pierwszy autor pracy, zadeklarował istotny udział w opracowaniu metody, implementację modeli, przeprowadzenie eksperymentów oraz analizę wyników.

Kontynuacją badań w zakresie prywatności danych, dla modeli autoregresyjnych, jest praca zatytułowana *„Privacy Attacks on Image AutoRegressive Models"*, która została opublikowana na renomowanej konferencji ICML. Generatywne modele autoregresyjne są interesujące, w kontekście identyfikacji danych źródłowych, z racji na ich rosnące znaczenie spowodowane wydajnością oraz własnościami zapamiętywania danych uczących. Są równocześnie bardziej podatne na wyciek prywatności w porównaniu z modelami dyfuzyjnymi. W identyfikacji danych wykorzystano wektor cech danych wejściowych bazujący na metodzie CLiD, jednak uwzględniający różnicę w odpowiedzi przy różnych warunkach. Dla wybranych modeli autoregresyjnych (MAR) autorzy zaproponowali cenne udoskonalenia, tj. adaptowalna maska binarna, stałe odstępy czasowe odszumiania, czy ograniczona wariancja szumu. Zagregowany sygnał z odpowiedzi MIA, dla danych treningowych, z pomocą testu statystycznego, pozwala sprawdzić czy model był trenowany za pomocą określonego (prywatnego)

zbioru danych. Wnioskowanie w metodzie zostało usprawnione poprzez wyeliminowanie liniowego klasyfikatora cech MIA, co można było zrobić stwierdzając konsekwentne różnice w odpowiedziach dla danych wykorzystywanych do treningu (*members*) i nie wykorzystywanych do treningu (*non-members*). Autorzy zauważyli również możliwość znacznej redukcji próbek identyfikujących dla większych modeli i konieczność większej liczby próbek, zapewniających istotność statystyczną wyników, dla małych modeli autoregresyjnych. Autorzy przeprowadzili też ciekawe eksperymenty ekstrakcji danych uczących z modeli regresyjnych bazując na ich własnościach zapamiętywania oraz stymulowania kontekstu w przestrzeni tokenów. Wyniki pokazują, że zaproponowane rozwiązanie pozwala zrekonstruować znacznie więcej danych (obrazów) uczących niż w rozwiązaniach referencyjnych i wyniki nie są obciążone fałszywie pozytywnie. Doktorant, będąc drugim autorem, zadeklarował istotny udział w tworzeniu mechanizmu audytowania danych i tworzeniu zbioru danych, implementację i przeprowadzenie eksperymentów oraz analizę wyników.

Podsumowując, w obszarze podjętej problematyki, Doktorant rzetelnie, kompleksowo i wieloaspektowo przeprowadził analizę problemów, zaproponował nowatorskie rozwiązania, zaimplementował algorytmy, zaprojektował eksperymenty i przeanalizował wyniki. W każdym z podjętych wątków był autorem przełomowych kontrybucji, wykazując się doskonałą wiedzą, intuicją i rzetelnością badawczą. Przedstawione rozwiązania, eksperymenty oraz wyniki są ciekawe i nowatorskie. Chociaż prace powstawały w zespole badaczy, to w każdej z rozważanych publikacji Doktorant był autorem lub współautorem kluczowych kontrybucji.

Na czym polega oryginalność rozprawy, co stanowi samodzielny i oryginalny dorobek Autora, jaka jest pozycja rozprawy w stosunku do stanu wiedzy, czy poziomu techniki reprezentowanych przez literaturę światową?

Oryginalność rozprawy doktorskiej leży przede wszystkim w zbiorze opracowanych metod, które nie tylko podejmują, analizują, ale również rozwiązują istotne zagadnienia bezpieczeństwa, zastosowania i niezawodności modeli generatywnych. W szczególności najważniejszymi, moim zdaniem, osiągnięciami są:

- opracowanie metody *ExpertSim* zapewniającej odpowiednią dywersyfikację trajektorii cząstek elementarnych, wizualizowanych przez symulator, w eksperymencie ALICE, przeprowadzanym w ośrodku CERN oraz potencjał metody do wizualizowania innych zjawisk fizycznych wysokich energii;

- opracowanie metody *B4B* aktywnego zabezpieczania enkoderów przed skopiowaniem bez zmniejszania użyteczności modelu dla zwykłych użytkowników;

- opracowanie metody CDI do efektywnego audytowania czy zadany zbiór danych był wykorzystany do trenowania modelu dyfuzyjnego;

- opracowanie nowej metody audytowania obrazowych modeli autoregresyjnych, weryfikującej czy dany zbiór danych był wykorzystany do nauki modelu oraz wykazanie, w jakim zakresie możliwe jest zreprodukowanie danych treningowych;

Chociaż wymienione osiągnięcia są, moim zdaniem, najważniejszymi z przedstawionych w dysertacji, to należy podkreślić, że każde z zaprezentowanych rozwiązań stanowi bezsprzecznie oryginalny wkład Doktoranta w rozwój dyscypliny Informatyka Techniczna i Telekomunikacja.

Czy Autor wykazał umiejętność poprawnego i przekonywującego przedstawienia uzyskanych przez siebie wyników (zwięzłość, jasność, poprawność redakcyjna rozprawy)? Jakie są słabe strony rozprawy i jej główne wady?

Praca stanowi bardzo szerokie studium analizy wybranych modeli generatywnych, zarówno w zakresie ich wykorzystania jak i audytu. Z racji przyjętej formy rozprawy, osiągnięcia zostały przedstawione bardzo treściwie, ale wyczerpująco – niezwykle cenne okazały się dodatki do publikacji, które uzasadniały szereg decyzji podjętych podczas opracowywania metod.

Praca jest zredagowana w języku angielskim i oprócz przywołania cyklu sześciu publikacji zawiera ciekawe wprowadzenie do całości jak i do poszczególnych części. Rozprawa zawiera też cenne podsumowania i uzupełnienia opisu badań. Z racji przyjętego kształtu rozprawy doktorskiej uważam, że bardzo trudno jest wskazać jakiekolwiek istotne uchybienia. Uwzględniając zakres prac badawczych ośmielam się stwierdzić, że takich uchybień, w prezentowanym materiale, po prostu nie ma.

Konkluzja

Uważam, że cele rozprawy doktorskiej, pomimo ich różnorodności, zostały w pełni zrealizowane. Doktorant przedstawił w rozprawie nowe badania z zakresu wykorzystania modeli generatywnych w symulacjach fizycznych wysokich energii, mechanizmy zabezpieczenia modeli przed skopiowaniem i metody audytowania modeli względem wykorzystania zbiorów danych w procesie ich nauki. Uzyskane przez Doktoranta wyniki uważam za oryginalne, wartościowe i ciekawe poznawczo. Zakres prowadzonych badań był szeroki, analiza opracowanych rozwiązań wszechstronna, a warsztat badawczy Doktoranta był właściwy. Tym samym rozprawa prezentuje wartościowe osiągnięcia naukowe w obszarze dyscypliny Informatyka Techniczna i Telekomunikacja oraz potwierdza umiejętność prowadzenia przez Doktoranta pracy naukowej.

Bez najmniejszej wątpliwości stwierdzam, że recenzowana rozprawa doktorska Pana mgr inż. Jana Dubińskiego spełnia z wyraźnym nadmiarem wymagania stawiane rozprawom doktorskim, przez obowiązującą ustawę. Wnoszę o jej przyjęcie i dopuszczenie rozprawy do publicznej obrony.

Wnoszę również o wyróżnienie rozprawy, gdyż zakres i jakość prezentowanych, oryginalnych rozwiązań naukowych przekracza zdecydowanie przeciętny poziom osiągnięć naukowych, uzyskiwanych przez doktorantów w większości znanych mi postępowań doktorskich. Na wyróżnienie zasługuje również dorobek publikacyjny, który prezentuje się spektakularnie i bezsprzecznie wymagał ogromnej wiedzy i determinacji Doktoranta.

Adam Wojciechowski