РЕЦЕНЗІЯ

на дисертаційну роботу
Мельниченка Артема Васильовича
На тему: «Методи та програмні засоби підвищення швидкодії моделей розпізнавання образів на основі машинного навчання»,
представлену на здобуття наукового ступеня доктора філософії
в галузі знань 12 – Інформаційні технології
за спеціальністю 121 – Інженерія програмного забезпечення

Актуальність теми дисертації

У епоху цифрової трансформації, нейронні мережі стали ключовою технологією інноваційних рішень в багатьох сферах, від медицини до автомобільної промисловості. Незважаючи на широкий спектр застосування та високі показники точності нейронних мереж, застосування даної технології висуває вимоги до обчислювальних ресурсів, значно ускладнюючи інтеграцію нейронних мереж у пристрої з обмеженими ресурсами. Тренування нейронних мереж вимагає використання дорогих і потужних графічних процесорів (GPU). Виконання нейронних мереж на кінцевих пристроях, що не містять графічних процесорів часто неможливе або ж спричиняє велику затримку, при цьому використання нейронних мереж в системах «розумного» дому, периферійних обчисленнях та мобільних застосунках могло б значно поширити доступність технології.

У відповідь на такі виклики та стрімкий розвиток нейронних мереж набули поширення методи обрізки і збільшення швидкодії нейронних мереж. Прунінг, квантизація, перенесення навчання (transfer learning) є ефективними методами, що дозволяють розширити застосування нейронних мереж і зменшити вимоги до обчислювальних ресурсів. Втім, часто обмеженням даних методів є втрата точності або ж складність в імплементації.

Метою дисертації є збільшення ефективності моделей нейронних мереж, а саме зменшення зменшення втрати точності при збільшенні швидкодії, після застосування методів оптимізації моделей глибинного навчання, створених для вирішення задач комп'ютерного зору.

Враховуючи вище сказане, актуальним є розробка методів та програмних засобів підвищення швидкодії моделей розпізнавання образів на основі машинного

навчання, що дозволить збільшити продуктивність нейронних мереж та надати технології більшого поширення і доступності.

Оцінка обґрунтованості наукових результатів дисертації, їх достовірності і новизни

Наукова новизна результатів дисертаційного дослідження полягає в наступному:

1. Удосконалено модель нейронної мережі для виявлення облич RetinaFace, яка на відміну від існуючих використовує метод прунінгу SNIP для оптимізації, що дозволяє використовувати розріджені матриці для зберігання і виконання мережі з метою подальшого удосконалення та збільшення швидкодії;

2. Удосконалено метод прунінгу SNIP для моделі виявлення облич RetinaFace, який на відміну від існуючих передбачає можливість виключення контекстних модулів з процесу прунінгу. Вдосконалений метод дозволяє досягти більшої точності при незмінній кількості виключених параметрів;

3. Вперше розроблено метод прунінгу перед навчанням для моделей архітектури трансформер, який на відміну від існуючих враховує важливість механізму «уваги». Використання розробленого методу дозволяє значно збільшити точність класифікації кінцевої моделі в порівнянні з методом SNIP;

4. Вперше розроблено архітектуру програмного забезпечення для моделювання та дослідження методів прунінгу перед навчанням нейронних мереж, яка на відміну від існуючих дозволяє приводити матриці вагових коефіцієнтів мережі до розрідженого формату, використовуючи запропонований механізм оцінки важливості вагів.

Достовірність наукових результатів забезпечується експериментальними дослідженнями, проведених із використанням сучасного обладнання. Наведені у роботі наукові положення, висновки та практичні рекомендації достатньо обґрунтовані, базуються на фактичних даних, які представлені у роботі в табличному та графічному вигляді. Наукові дослідження були виконані здобувачем на кафедрі інженерії програмного забезпечення в енергетиці «КПІ ім. Ігоря Сікорського» в рамках НДР «Методи і алгоритми оптимізації розпізнавання образів на основі методів машинного навчання» Державний реєстраційний номер: № 0121U109207, що виконувались в Національному технічному університеті України «Київський політехнічний інститут імені Ігоря Сікорського».

Результати досліджень прийняті до впровадження в Товаристві з обмеженою відповідальністю «ВОТЧЕД» (Акт від 20.10.2023р.), що підтверджено відповідним Актом впровадження.

Отже, в дисертаційній роботі поставлене наукове завдання розробки науково-методичного апарату та ПЗ для моделювання і оптимізації процесу навчання та застосування нейронних мереж для вирішення задач розпізнавання образів з метою підвищення швидкодії моделей розпізнавання образів виконано повністю, здобувач повною мірою оволодів методологією наукової діяльності.

Оцінка змісту дисертації, її завершеність та дотримання принципів

За своїм змістом, дисертація Мельниченка А.В. повністю відповідає Стандарту вищої освіти зі спеціальності 121 – Інженерія програмного забезпечення та напрямкам дослідження відповідно до освітньої програми Інженерія програмного забезпечення.

Дисертаційна робота є завершеною науковою працею і свідчить про наявність особистого внеску здобувача у науковий напрям розробки програмного забезпечення для «Глибоких нейронних мереж».

Розглянувши звіт подібності за результатами перевірки дисертаційної роботи на текстові співпадіння, можна зробити висновок, що дисертаційна робота Мельниченка Артема Васильовича є результатом самостійних досліджень здобувача і не містить елементів фальсифікації, компіляції, фабрикації, плагіату та запозичень. Використані ідеї, результати і тексти інших авторів мають належні посилання на відповідне джерело. Принципи академічної доброчесності не були порушені.

Мова та стиль викладення результатів академічної доброчесності

Дисертація написана українською мовою. Дисертація Мельниченка А.В. є структурованою, матеріал викладено послідовно, у науковому стилі мовлення із дотриманням сучасної загальноприйнятої термінології. Дисертація складається з вступу, чотирьох розділів, висновків, списку використаних джерел та додатків. Загальний обсяг дисертації становить 179 сторінок, з яких 145 сторінок основного тексту, 3 додатки на 19 сторінках, та містить 53 рисунки, 24 формул, 9 таблиць.

У вступі розглядається актуальність дисертаційного дослідження, зв'язок роботи з науковими програмами, планами та темами. Формулюється мета, завдання, об'єкт та предмет дослідження, наукова та практична новизна отриманих

результатів. Приводяться відомості про особистий внесок здобувача та апробацію результатів дисертації.

У першому розділі зроблено огляд та аналіз сучасних підходів до вирішення задач розпізнавання образів, методів обрізки та методів збільшення швидкодії нейронних мереж. Обґрунтовується актуальність обраного класу методів як одного з обрізки глибоких нейронних мереж.

У другому розділі викладено опис архітектури нейронної мережі RetinaFace для виявлення облич, метод прунінгу перед навчанням SNIP та результати експериментів із застосування методу для прунінгу даної мережі з подальшою модифікацією. Для розробки модифікації методу зроблено припущення, що контекстні шари архітектури є більш важливими, тому варто надати їм більшу вагу при обрахунку критерію важливості. Також, проаналізовано альтернативний метод оптимізації моделей, що являє собою додавання до архітектури моделі додаткових виходів задля кращої здатності узагальнення знань, який не показав збільшення точності чи швидкодії.

У третьому розділі викладено опис архітектури нейронної мережі «трансформер» для класифікації зображень, розроблений метод прунінгу з використанням критеріїв важливості вагів, що враховує механізм самоуваги в даній архітектурі нейронної мережі, та результати експериментів із застосування модифікованого методу для прунінгу даної мережі. Також, було запропоновано підхід для підвищення швидкодії при використанні модифікованого методу, який використовує розріджені матриці формату 2:4 для пришвидшення обчислень.

У четвертому розділі описано розроблену архітектуру програмної системи для дослідження ефективності методів прунінгу нейронних мереж перед навчанням з використанням хмарних технологій та мікросервісного підходу.

У Додатках наводяться акти впровадження та програмний код розроблених компонентів для прунінгу та проведення експериментів.

Дисертаційна робота оформлена відповідно до вимог наказу МОН України від 12 січня 2017 р. №40 «Про затвердження вимог до оформлення дисертації».

Оприлюднення результатів дисертаційної роботи

Наукові результати дисертації висвітлені у 3 наукових публікаціях здобувача, серед яких: 3 статті у фахових наукових виданнях України категорії «Б». Також результати дисертації були апробовані на 7 наукових фахових конференціях.

Усі публікації здобувача мають високий науковий рівень, в них достатньо повно описані головні наукові здобутки, представлені в дисертації. Порушення принципів академічної достовірності не виявлені. Особистий внесок здобувача до всіх наукових публікацій, опублікованих у співавторстві, є вагомим.

Таким чином, наукові результати описані в дисертаційній роботі повністю висвітлені у наукових публікаціях здобувача.

Недоліки та зауваження до дисертаційної роботи

1. Здобувач наводить в актуальності теми виклики в застосування нейронних мереж на мобільних пристроях та у сфері IoT, при цьому проведені експерименти не включають замірів на даних пристроях. Було б доцільно включити заміри швидкодії та енерговитрат на таких пристроях;
2. Розроблені методи, описані у розділах 2 і 3, базуються на розрахунку критерію та порівнюються лише з методом SNIP. Було б доцільно розширити перелік методів для порівняння;
3. Розроблені методи враховують унікальні особливості наведених архітектур, втім не зрозуміло, чи можна поширити застосування даних методів на інші мережі;
4. Вибір класу методів прунінгу перед навчанням потребує додаткового обґрунтування. Незрозуміло, чому автор фокусується саме на прунінгу перед навчанням, і не проводить порівняння з методами прунінгу після навчання;
5. У четвертому розділі описано розроблену архітектуру для хмарного середовища Azure та використано фреймворк PyTorch, при цьому не було проведено порівняння з іншими фреймворками і хмарними середовищами, такими як Tensorflow та AWS/Google Cloud.

Вважаю, що висловлені зауваження не є визначальними і не зменшують загальну наукову новизну та практичну значимість результатів та не впливають на позитивну оцінку дисертаційної роботи.

Висновок про дисертаційну роботу

Вважаю, що дисертаційна робота здобувача ступеня доктора філософії Мельниченка Артема Васильовича на тему «Методи та програмні засоби підвищення швидкодії моделей розпізнавання образів на основі машинного навчання» виконана на високому науковому рівні, не порушує принципів академічної доброчесності та є закінченим науковим дослідженням, сукупність

теоретичних та практичних результатів якого розв'язує наукове завдання, що має істотне значення для галузі знань Інформаційні технології. Дисертаційна робота за актуальністю, практичною цінністю та науковою новизною повністю відповідає вимогам чинного законодавства України, що передбачені в п.6 - 9 «Порядку присудження ступеня доктора філософії та скасування рішення разової спеціалізованої вченої ради закладу вищої освіти, наукової установи про присудження ступеня доктора філософії», затвердженого Постановою Кабінету Міністрів України від 12 січня 2022 р. №44.

Здобувач Мельниченко Артем Васильович заслуговує на присудження ступеня доктора філософії в галузі знань 12 - Інформаційні технології за спеціальністю 121 – Інженерія програмного забезпечення.

Рецензент:
В.о. завідувача кафедри інженерії програмного забезпечення в енергетиці
навчально-наукового інституту атомної та теплової енергетики
Національного технічного університету України
«Київський політехнічний інститут імені Ігоря Сікорського»

доктор технічних наук, професор Олександр КОВАЛЬ

Підпис гр.
ЗАСВІДЧУЮ
Відділ кадрів та діловодства
підпис пр-ще

М.П.

Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського» Відділ кадрів і архівної справи № 6

«6» ЧЕРВНЯ 2024 року