ВІДГУК

офіційного опонента на дисертаційну роботу

Тарана Владислава Ігоровича

на тему «Метод адаптації глибоких нейронних мереж до апаратного забезпечення зі спеціалізованою архітектурою»,

представлену на здобуття ступеня доктора філософії

в галузі знань 12 - Інформаційні технології

за спеціальністю 123 - Комп'ютерна інженерія

Актуальність теми дисертації. Сучасний прогрес у розробці і застосуванні засобів штучного інтелекту та машинного навчання революціонізує багато сфер людської діяльності, зокрема створюючи основу для нової промислової революції (Індустрія 4.0). В цьому контексті актуальним і перспективним є напрямок впровадження результатів машинного навчання у спеціалізованих обчислювальних архітектурах, що дають можливість ефективного застосування моделей штучного інтелекту на мобільних пристроях (смартфонах, планшетах, носимій електроніці), у вбудованих системах (Embedded Systems – промисловому обладнанні з числовим управлінням, в інтелектуальних системах безпеки, моніторингу здоров'я, автономного водіння тощо), у розподілених системах периферійних обчислень (Edge Computing). Обмеження доступної пам'яті та обчислювальної продуктивності разом з вимогами енергоефективності та компактності визначають специфічні вимоги до застосування в таких системах нейронних мереж глибокого навчання – найбільш успішного сучасного алгоритму штучного інтелекту. Звідки випливає актуальність розробки методів адаптації нейромережевих моделей для спеціалізованих обчислювальних архітектур шляхом балансування їх складності та якості роботи. Я би відмітив ще одну галузь, в якій така адаптація є надзвичайно корисною: це розробка хмарних інтелектуальних систем. Застосування спрощених моделей може забезпечити суттєву економію на ціні орендованих ресурсів.

Оцінка обґрунтованості наукових результатів дисертації, їх достовірності та новизни. Головний науковий результат роботи полягає у розробленому здобувачем новому комплексному методі адаптації глибоких нейронних мереж для спеціалізованих обчислювальних архітектур, що інтегрує наступні складові:

- новий метод адаптивного ітеративного прунінгу для зменшення розміру моделей, що відрізняється від відомих аналогів адаптивним балансуванням точності та швидкості через регуляцію гіперпараметрів;

- метод підвищення ефективності (тобто продуктивності) процесу обробки даних шляхом зменшення накладних витрат за рахунок підбору параметрів таких, як розмір пакету даних, для певної апаратури;
- метод підвищення ефективності обчислювальної інфраструктури за рахунок підбору якісних параметрів (типу операційної системи, інтерфейсу підключення пристрою, бібліотеки тощо).

Достовірність отриманих наукових результатів дослідження підтверджена наведеними в тезах дисертації результатами експериментів і прикладами застосування адаптації до реальних моделей, зокрема для діагностики рентгенівських знімків, та ґрунтується на наведеному огляді сучасних технологій модифікації глибоких нейронних мереж для підвищення продуктивності обробки даних. Подані у роботі наукові положення, висновки та практичні рекомендації повністю обґрунтовані, базуються на фактичних даних, що представлені в тезах у табличному і графічному вигляді.

Результати одержані здобувачем пройшли ретельну фахову експертизу при рецензуванні відповідних статей в журналах з високим імпакт-фактором і тез міжнародних конференцій, що теж свідчить про їх оригінальність, обґрунтованість та достовірність.

Теоретичне та практичне значення роботи Тарана В.І. полягає у створенні засад ефективного застосовувати моделей нейронних мереж глибокого навчання на різноманітному апаратному забезпеченні. Розроблений комплексному метод адаптації нейронних мереж до спеціалізованих обчислювальних архітектур вдало поєднує методи прунінгу, квантизації, видалення зайвих класів об’єктів, підвищення продуктивності процесу обробки даних певною моделлю з паралельним підвищенням ефективності програмно-апаратної складової цільової інфраструктури. Варто відмітити і такий суто прикладний результат роботи, як розроблений програмний компонент діагностики легеневих аномалій за даними рентгенівських знімків для спеціалізованого прискорювача Coral Edge TPU USB.

Отже, можна впевнено стверджувати, що поставлене завдання дослідження виконано здобувачем на високому науковому рівні. Здобувач продемонстрував впевнене володіння методологією наукової діяльності.

Оцінка змісту дисертації, її завершеність та дотримання принципів академічної доброчесності. Дисертаційна робота Тарана В.І. є завершеною науковою працею, що за своїм змістом повністю відповідає Стандарту вищої освіти зі спеціальності «123 - Комп’ютерна інженерія» за напрямком досліджень відповідним до освітньої програми «Комп’ютерна інженерія». Тези дисертаційної роботи підтверджують особистий внесок здобувача у науковий напрям «Глибокі нейронні мережі».

Аналіз тексту дисертації, зокрема, звіт подібності за результатами автоматичної перевірки на текстові співпадіння свідчить, що дисертаційна робота 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.

Тези дисертації доводять, що поставлене завдання дослідження виконано здобувачем повністю, самостійно, з додержанням академічної доброчесності.

Мова та стиль викладення результатів. Дисертаційна робота Тарана В.І. є добре структурованою. Розділи та підрозділи тез адекватно розподілені, відображають суть проведених наукових досліджень. Викладення роботи є послідовним, чітко обґрунтовує отримані результати. Дисертація написана фаховою українською мовою.

Дисертація складається з вступу, чотирьох розділів, висновків, переліку посилань та додатків. Загальний обсяг дисертації становить 152 сторінки.

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 зроблений аналітично-порівняльний огляд сучасних спеціалізованих обчислювальних архітектур для застосування нейронних мереж глибокого навчання, сучасних методів прунінгу та квантизації для зменшення розміру нейромережевих моделей. Описані популярні нейронні мережи глибокого навчання та тестові колекції даних.

У другому розділі розглянуто складові адаптації глибоких нейронних мереж для спеціалізованих прискорювачів. Проаналізовано обчислювальну складність застосування моделей і фактори апаратно-програмного середовища. Запропоновані метрики оцінки продуктивності та ефективності адаптації.

У третьому розділі детально описано запропонований здобувачем комплексний метод адаптації разом з складовими методами. Наведено приклад його застосування для адаптації відомої мережі інтелектуального аналізу зображень ResNet50 під тензорний прискорювач Coral Edge TPU та описано програму діагностики легеневих аномалій за рентгенівськими знімками, що використовує адаптовану модель.

У четвертому розділі експериментально проаналізовано результати застосування комплексного методу адаптації глибоких нейронних мереж та його складових. Оцінено ступінь впливу окремих факторів продуктивності, досліджено продуктивність розробленого медичного додатку.

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 12 наукових публікаціях здобувача, серед яких: 6 статей у періодичних наукових виданнях іноземних держав проіндексованих у базах Scopus / WoS, з них 1 стаття - у виданні, віднесеному до першого квартилю (Q1) відповідно до класифікації SCImago Journal and Country Rank / Journal Citation Reports, 6 тез у збірниках доповідей фахових наукових конференцій.

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

Недоліки та зауваження до дисертаційної роботи. Констатуючи високий науковий рівень виконаного дослідження, слід зауважити, що автору не вдалося уникнути деяких недоліків.

1. Природною альтернативою адаптації складної моделі до менш продуктивної системи є конструювання і тренування простішої моделі (з меншою кількістю, або меншим розміром шарів). Порівняння адаптованих моделей з початковими варто було б доповнити порівняннями з іншими тренованими моделями того ж розміру, як адаптовані. Це було б корисно як для демонстрації переваги адаптації, так і для більш інтуїтивної оцінки якості її результату в конкретних експериментах.

2. Запропонований в роботі комплексний метод адаптації глибоких нейронних мереж для спеціалізованих обчислювальних архітектур складається з 4 кроків, останні два з яких можна інтерпретувати як підбор кількісних і якісних параметрів середовища виконання програми. Не зрозуміло, чому підбір кількісних параметрів передує підбору якісних, а не навпаки. Можна очікувати що ці параметри взаємозалежні (тобто, наприклад, зміна операційної системи, чи апаратного інтерфейсу може бути підставою для вибору іншого розміру пакету внаслідок іншої латентності та швидкості передачі даних на пристрій). Можливо останні два кроки треба виконувати циклічно, або взагалі об’єднати в один?

3. Як доповнення до попереднього зауваження: було б доцільно описати, як організувати (а можливо і автоматизувати) багатофакторну оптимізацію параметрів середовища, зокрема, для периферійних обчислень у динамічному і гетерогенному апаратному середовищі.

4. З практичної точки зору важливою є ситуація, коли модель не може бути спрощена до такого ступеню, щоб вміститись у певну обчислювальну архітектуру. (Або таке спрощення призведе до неприйнятного падіння якості обробки даних.) Я не знайшов у роботі умов і рецептів для такої ситуації.

5. Парною до задачі адаптації нейромережної моделі під спеціалізовану архітектуру є задача підбору обчислювальної архітектури під модель. Було б доцільно розглянути проблему і під таким кутом.

Висловлені зауваження не є визначальними, не заперечують наукову новизну і практичну значимість отриманих результатів та не впливають на загальну позитивну оцінку дисертаційної роботи.

Висновок про дисертаційну роботу. Вважаю, що дисертаційна робота здобувача ступеня доктора філософії Тарана Владислава Ігоровича на тему «Метод адаптації глибоких нейронних мереж до апаратного забезпечення зі спеціалізованою архітектурою» 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 Таран Владислав Ігорович заслуговує на присудження ступеня доктора філософії в галузі знань «12 - Інформаційні технології» за спеціальністю «123 - Комп'ютерна інженерія».

Офіційний опонент:
Завідувач відділу
автоматизації програмування № 145,
Інститут кібернетики
ім. В.М. Глушкова НАН України,
доктор фізико-математичних наук,
старший науковий співробітник Вадим ТУЛЬЧИНСЬКИЙ

М.П. «9» січня 2023 року

Національна академія наук України * м. Київ * Інститут кібернетики імені В.М.Глушкова ідентифікаційний код 05417176