

## **РЕЦЕНЗІЯ**

на дисертаційну роботу **Тарана Владислава Ігоровича**

на тему «Метод адаптації глибоких нейронних мереж до апаратного забезпечення зі спеціалізованою архітектурою», подану на здобуття ступеня доктора філософії в

галузі знань 12 Інформаційні технології

за спеціальністю 123 Комп'ютерна інженерія

### **Актуальність теми дисертації**

Значного поширення глибокі нейронні мережі отримали через стрімкий прогрес високопродуктивних обчислень та спеціалізованих обчислювальних архітектур. За рахунок збільшення кількості шарів мережі та додавання спеціалізованих блоків сучасні нейронні мережі досягли високих показників точності у задачах комп'ютерного зору, детекції, розпізнавання, обробки мови тощо. Зі зростанням складності архітектури нейронних мереж відповідно зростають вимоги до апаратних засобів, на яких виконується обробка даних. В цьому контексті важливого значення набувають прискорювачі глибоких нейронних мереж зі спеціалізованою архітектурою, основною складовою яких є спеціалізовані обчислювальні модулі Tensor Cores. Оскільки пам'ять таких прискорювачів є значно меншою у порівнянні із звичайними GPU, виникає необхідність або адаптації нейронних мереж шляхом конвертації архітектури мережі у формат, підтримуваний конкретним пристроєм, або зменшення розміру мережі. Особливо використання подібних прискорювачів є актуальним у сфері Edge Computing, де наявні набагато більші обмеження на обчислювальні ресурси та вимоги високої енергоефективності. Враховуючи зазначені обмеження спеціалізованих прискорювачів, доцільно є розробка методу адаптації глибоких нейронних мереж для спеціалізованих обчислювальних архітектур, що забезпечує обробку даних з високою продуктивністю без втрати точності розпізнавання.

### **Оцінка обґрунтованості наукових результатів дисертації, їх достовірності та новизни**

Наукова новизна результатів дисертаційного дослідження полягає в наступному:

– **вперше** запропоновано метод адаптації глибоких нейронних мереж, що надає можливість застосовувати комплексний підхід в адаптації і, на відміну від існуючих підходів, дозволяє підготувати модель глибокої нейронної мережі для прискорювачів зі спеціалізованою архітектурою, включно із її використанням на рівні периферійних обчислень;

– **вперше** розроблено метод підвищення ефективності процесу обробки даних глибокими нейронними мережами на спеціалізованих обчислювальних архітектурах, що враховує накладні витрати ініціалізації та передачі даних на спеціалізованих прискорювачах та, на відміну від існуючих підходів, дозволяє визначити гіперпараметри мережі, щоб підвищити продуктивність роботи мережі та зменшити час виконання ітерації обробки даних;

– **вперше** розроблено метод підвищення ефективності інфраструктури для обробки даних, що включає в себе зміну програмної та апаратної конфігурації і, на відміну від існуючих підходів, дозволяє зменшити час виконання ітерації обробки даних глибокою нейронною мережею на цільовій системі;

– **набув подальшого розвитку** метод прунінгу за рахунок ітеративного зменшення розміру моделі глибокої нейронної мережі та адаптивної зміни гіперпараметрів мережі, що дозволяє компенсувати втрати точності після кожної ітерації прунінгу та збільшити продуктивність обробки даних глибокою нейронною мережею.

Наукове дослідження було виконано здобувачем на кафедрі обчислювальної техніки КПІ ім. Ігоря Сікорського згідно затвердженого плану наукової роботи кафедри, що враховує розпорядження Кабінету Міністрів України від 2 грудня 2020 р. № 1556-р про схвалення Концепції розвитку штучного інтелекту в Україні. Запропоновані у дисертації методи використані у науково-дослідній роботі Національного фонду досліджень України «Наука для безпеки людини та суспільства» – проект «Платформа штучного інтелекту для дистанційного автоматизованого виявлення та діагностики захворювань людини», реєстраційний номер проекту 2020.01/0490.

### **Оцінка змісту дисертації, її завершеність та дотримання принципів академічної добродетелі**

За своїм змістом дисертація здобувача Тарана В.І. відповідає стандарту вищої освіти третього (освітньо-наукового) рівня, галузі знань 12 Інформаційні технології, спеціальністі 123 Комп’ютерна інженерія та освітньо-науковій програмі третього (освітньо-наукового) рівня «Комп’ютерна інженерія».

Дисертація є завершеною науковою працею і свідчить про наявність особистого внеску здобувача у науковий напрям “Глибокі нейронні мережі”. У дисертаційному дослідженні здобувачем виконано наукове завдання адаптації глибоких нейронних мереж до апаратного забезпечення зі спеціалізованою архітектурою, здобувач повною мірою оволодів методологією наукової діяльності. У дисертації не виявлено елементів фальсифікації, компіляції, фабрикації, plagiatu, запозичень тощо. Використані ідеї, результати і тексти інших авторів мають посилання на відповідні джерела.

### **Мова та стиль викладення результатів**

Дисертація написана українською мовою, складається з анотації, написаної українською та англійською мовами, вступу, 4 розділів основної частини, висновків, списку літератури із 118 джерел та додатків. Загальний обсяг дисертації становить 152 сторінки, з яких 103 сторінки основного тексту, 2 додатки на 24 сторінках, та містить 58 рисунків, 24 формул, 6 таблиць. Оформлення дисертації відповідає вимогам наказу МОН України від 12 січня 2017 р. № 40 «Про затвердження вимог до оформлення дисертації».

У вступі обґрунтована актуальність теми, зв’язок роботи з науковими програмами, планами та темами, сформульовані мета та завдання, об’єкт, предмет

та методи дослідження, наукова та практична новизна отриманих результатів, наведені відомості з апробації та публікації результатів дисертації із зазначенням особистого внеску здобувача.

У першому розділі представлено огляд сучасних технологій зменшення глибоких нейронних мереж для підвищення продуктивності та ефективності обробки даних. У наступному розділі розглянуто особливості процесу адаптації глибоких нейронних мереж із застосуванням спеціалізованих прискорювачів. Зроблено висновок про необхідність врахування усіх факторів, що впливають на продуктивність та ефективність обробки даних нейронними мережами, в комплексі.

У третьому розділі описано розроблений здобувачем метод адаптації глибоких нейронних мереж до апаратного забезпечення зі спеціалізованою архітектурою, що включає адаптивний ітеративний прунінг для зменшення розміру моделей нейронних мереж, підвищення ефективності процесу обробки даних нейронними мережами на спеціалізованих обчислювальних архітектурах та підвищення ефективності обчислювальної інфраструктури для глибоких нейронних мереж. Слід відмітити, що запропонований метод досягає прискорення обчислень з обробки даних як за рахунок скорочення обсягу обчислювальних операцій, так і за рахунок використання спеціалізованих пристройів. Саме таке комбінування методів дає змогу досягти високої продуктивності та достатньої ефективності.

У четвертому розділі доведено експериментально ефективність застосування методу адаптації глибоких нейронних мереж на прикладі обробки рентген зображень.

### **Оприлюднення результатів дисертаційної роботи**

Наукові результати дисертації повністю висвітлені у 8 наукових публікаціях здобувача, серед яких 6 статей у періодичних наукових виданнях іноземних держав, проіндексованих у базі Scopus, у тому числі 1 стаття у виданні, віднесеному до першого квартилю (Q1) відповідно до класифікації SCImago Journal and Country Rank та 4 статті в двох номерах журналів, 2 публікації у матеріалах міжнародних наукових конференцій. Результати дисертації були апробовані на 4 міжнародних наукових конференціях.

### **Недоліки та зауваження до дисертаційної роботи**

1. Розділ 2 повинен містити теоретичні розробки з дисертаційного дослідження такі, як формалізований опис задачі, певні теоретичні факти, на яких ґрунтуються робота, у тому числі виведені самостійно теоретичні твердження. Замість цього розділ 2 присвячений обґрунтуванню необхідності оптимізації, як і розділ 1.

2. Відсутня математична формалізація запропонованого методу. Не сформульовано математично критерій ефективності.

3. Формальні вирази складені недостатньо чітко. Наприклад, у виразах (3.6)-(3.8) використовується за словами автора «невідома функція, яка може бути визначена емпірично». Однак приклад такого визначення автор в роботі не

наводить. Складність такого визначення може бути завелика для практичного застосування.

4. Застосування запропонованих методів, дійсно, буде корисним для комп'ютерного зору, як це вказано на стор. 44, проте приклад застосування методу та його експериментальне дослідження наведено для розпізнавання хвороби легенів, де швидкодія не настільки критична.

5. Щодо застосування в медичній діагностиці, то питання точності, напевно, є більш важливим, ніж швидкодія. У зв'язку з цим в практичній реалізації запропонованого методу слід дотримуватись чітко сформульованих вимог щодо точності, ефективності та продуктивності.

6. Велике значення прискорення, яке отримано експериментально, може бути обумовлено надто великою архітектурою нейромережі, з якою виконується порівняння.

Вищезазначені зауваження не зменшують загальну наукову новизну та практичну значимість дисертаційного дослідження, і не впливають на 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 добросовісності та є завершеним науковим дослідженням. Дисертаційна робота за актуальністю, практичною цінністю та науковою новизною повністю відповідає вимогам чинного законодавства України, що передбачені пп. 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.

### **Рецензент:**

Професор кафедри інформатики  
та програмної інженерії,  
КПІ ім. Ігоря Сікорського,  
доктор технічних наук, професор

