ВІДГУК

Офіційного опонента 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 контексті стрімкого розвитку технологій, методи машинного навчання, зокрема глибокі нейронні мережі, стають ключовим інструментом у сфері інформаційних технологій. Глибокі архітектури стали ефективним рішенням в широкому спектрі завдань, від розпізнавання образів до обробки природної мови та забезпечення безпеки інформаційних систем, що робить їх незамінним елементом сучасних технологічних рішень. Втім, середовище виконання нейронних мереж часто обмежується їх високою обчислювальною складністю та великою потребою в ресурсах. Широкий клас застосувань нейронних мереж потребує певної пропускної здатності, що також зумовлює потребу в збільшенні кількості ресурсів або ж зменшенні обчислювальної складності моделі. Наведені обмеження часто слугують блокуючим фактором при запуску глибоких нейронних мереж на пристроях Інтернету речей (IoT) та мобільних платформах, де ресурси зазвичай обмежені.

Подальше розширення можливостей використання нейронних мереж в умовах обмежених обчислювальних ресурсів вимагає значної оптимізації. Техніки квантування, прунінг та дистиляція знань є основними інструментами для обрізки та збільшення швидкодії нейронних мереж, що дозволяють значно скоротити кількість параметрів і обчислювальні вимоги без істотної втрати точності. Ці

методи мають велике значення для розгортання складних мереж на пристроях з обмеженими ресурсами. Тому тема розробки методів та програмних засобів підвищення швидкодії моделей розпізнавання образів на основі машинного навчання є актуальною.

Оцінка обґрунтованості наукових результатів дисертації, їх достовірності і новизни

Головний науковий результат роботи полягає у розробленій здобувачем архітектурі програмного забезпечення і методах для обрізки і збільшення швидкодії глибоких нейронних мереж. Даний результати включає наступну наукову новизну:

- 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;

- 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;

- 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;

- 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 отриманих в роботі наукових результатів та висновків забезпечується використанням класичних методів досліджень та підтверджується експериментально. В цілому сукупність результатів досліджень є інформативною і узгоджується з сучасними теоретичними положеннями.

Отже, в дисертаційній роботі поставлене наукове завдання виконане повністю, здобувач повною мірою оволодів методологією наукової діяльності.

Оцінка змісту дисертації, її завершеність та дотримання принципів академічної доброчесності

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 «Глибокі нейронні мережі».

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.

Мова та стиль викладення результатів

Дисертація написана українською мовою. Дисертація Мельниченка А.В. є добре структурованою. Оформлення роботи є послідовним та чітко підкреслює найважливіші результати, що були отримані. 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.

Перший розділ містить аналітичний огляд сучасних підходів до вирішення задач розпізнавання образів та методів підвищення продуктивності нейронних мереж. Обґрунтовується актуальність прунінгу як одного з методів підвищення швидкодії глибоких нейронних мереж.

Другий розділ містить опис провідної архітектури для виявлення обличь RetinaFace, опис набору даних для тренування WIDERFace та актуального методу прунінгу SNIP, що базується на критерії важливості для вагів. Приводяться ілюстрації набору даних, кількісні та якісні характеристики, що обґрунтовують вибір набору для експериментів. Обґрунтовується вибір архітектури та особливості її використання. Приводиться запропонована модифікація методу SNIP, що передбачає виключення окремих шарів нейронної мережі з обрахунку критерію. Описано експеримент та наведено дані, що демонструють порівняння обох методів при застосуванні для тренування моделі архітектури RetinaFace. Приводиться дані експерименту з сіамськими нейронними мережами та додатковими виходами, що не показав збільшення точності чи швидкодії.

Третій розділ містить детальний опис роботи нейронних мереж архітектури «трансформер» і обґрунтування проблеми обчислювальної складності таких мереж. Описано модифікацію даної архітектури для вирішення задач розпізнавання образів (ViT). Приведений опис розробленого алгоритму розрахунку критерію важливості вагів для архітектури трансформер, що враховує оцінки механізму уваги, а також порівняльний експеримент розробленого методу та методу SNIP на наборі даних для класифікації хвороб листя рослин. Також розділ містить аналіз використання прунінгу для збільшення швидкодії та зменшення обчислювальної складності,

використання розріджених матриць для прискорення виконання нейромереж та поєднання результатів прунінгу з використанням розріджених матриць.

Четвертий розділ містить опис архітектури розробленого програмного забезпечення для використання розроблених методів та збільшення швидкодії нейронних мереж. У розділі описано використані засоби розробки, фреймворки та бібліотеки для тренування нейронних мереж, надано розроблену сервісну архітектуру програмного забезпечення, описано реалізацію сервісів та застосування даної архітектури для тренування і збільшення швидкодії нейронних мереж з використанням хмарного середовища Azure.

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. Вимоги до програмного забезпечення, описані в першому розділі, потребують більш чіткого обґрунтування.

2. Наведено використання методів прунінгу для обрізки мережі архітектури RetinaFace, що базується на згорткових нейронних мережах, і порівняння

модифікованого методу з існуючим методом SNIP. Доцільно було б провести порівняння і для інших методів прунінгу, як приклад, прунінг фільтрів згорткових шарів.

3. Наведені в першому розділі методи збільшення швидкодії було б доцільно застосувати для оцінки швидкодії отриманої в другому розділі моделі.

4. У третьому розділі описано модифікацію ViT під назвою Linformer, проте не наводиться кількісних характеристик щодо впливу використання модифікації у порівнянні зі звичайним ViT.

5. Описаний в третьому розділі набір даних є меншим, ніж в другому, і задача змінюється з виявлення облич на класифікацію зображень. Хоча обидві задачі є частковими задачами розпізнавання образів, було б доцільно порівняти натреновані мережі на одному й тому самому наборі даних.

Вважаю, що висловлені зауваження не є визначальними і не зменшують загальну наукову новизну та практичну значимість результатів і 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, є закінченим науковим дослідженням, не порушує принципів академічної доброчесності, 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.

Офіційний опонент:
Завідувач кафедри Інженерії програмного забезпечення автоматизованих систем
Державного університету інформаційно-телекомунікаційних технологій

доктор технічних наук, професор Каміла СТОРЧАК

Підпис професора Сторчак К.П. з а с в і д ч у ю:

Учений секретар
Державного університету
інформаційно-комунікаційних технологій Галина ЄНЧЕВА

«6» ЧЕРВНЯ 2024 р.

М.П.