Голові
разової спеціалізованої вченої ради
Інституту математики НАН України
доктору фізико-математичних наук,
заступнику директора з наукових питань
Інституту математики НАН України
Василику Віталію Богдановичу

ВІДГУК

офіційного опонента на дисертаційну роботу
ШАМРАЯ Максима Борисовича на тему:
«Low-rank approximations, perturbation bounds, and their role in model compression»
(«Низькорангові наближення, оцінки збурень та їх роль у компресії моделей»),
подану на здобуття ступеня доктора філософії
у галузі знань 11 Математика та статистика
за спеціальністю 113 Прикладна математика

У дисертаційній роботі Максима Борисовича Шамрая представлені нові наукові результати, які стосуються побудови строго обґрунтованих підходів до зменшення розмірності моделей машинного навчання. Запропоновані методи спираються на поєднання низькорангових апроксимацій, інструментарію теорії збурень та аналізу структурних трансформацій параметрів моделей, що забезпечує їхню теоретичну вивіреність. Дисертація присвячена розв'язанню важливої проблеми сучасної прикладної математики, що полягає у формуванні узагальненого підходу до компресії великомасштабних моделей із контрольованими оцінками похибки та збереженням якості розв'язання відповідних задач. Такий підхід має принципове значення для математичного моделювання нейронних мереж і їх практичного застосування. Відповідно, ***актуальність обраної тематики*** не викликає сумніву. Вона зумовлена стрімким зростанням складності та розмірності сучасних нейронних архітектур, що, у свою чергу, призводить до суттєвих обмежень щодо обчислювальних ресурсів, пам'яті та енергоспоживання. На цьому тлі, існуючі підходи до компресії здебільшого базуються на евристичних міркуваннях і не забезпечують достатнього рівня теоретичних гарантій, що підкреслює доцільність і своєчасність проведеного дослідження.

Наукова новизна дисертаційної роботи полягає у розробленні цілісного підходу до задачі компресії моделей машинного навчання, який базується на системному поєднанні різних математичних методів у межах єдиної теоретично

обґрунтованої конструкції. Зменшення розмірності інтерпретується як керований процес трансформації параметрів, для якого забезпечується можливість формалізованого оцінювання впливу на точність відтворення та якість розв’язання задач. Автором обґрунтовано доцільність спільного розгляду методів низькорангового наближення, аналізу змін сингулярного спектра та структурного спрощення моделей, що дозволяє отримати більш повне уявлення про природу похибок, які виникають у процесі компресії. У результаті сформовано підхід, у межах якого характеристики точності не є лише емпіричними оцінками, а визначаються як складова частина відповідних алгоритмічних процедур. Отримані результати мають ***практичну цінність***, оскільки створюють підґрунтя для застосування компресованих моделей у задачах, чутливих до зміни їх поведінки. Запропоновані методи забезпечують більш передбачуваний характер таких змін, що є важливим у контексті використання моделей у середовищах з обмеженими обчислювальними ресурсами.

Основними науковими результатами даної дисертаційної роботи є:

- Доведено узагальнені співвідношення для збурень сингулярного спектра конкатенованих матриць, що дозволяють кількісно оцінювати вплив локальних змін блокової структури на їх спектральні властивості.
- Запропоновано математичну постановку задачі групування матриць із урахуванням обмеження на допустиму похибку апроксимації, внаслідок чого побудовано підґрунтя для переходу від інтуїтивних схем до формалізованих процедур із наперед заданими характеристиками.
- Отримано як глобальні, так і покрокові оцінки похибки реконструкції, що забезпечують можливість поетапного об’єднання матриць із контрольованим накопиченням похибки та гарантіями якості результату.
- Розроблено декілька підходів до кластеризації в задачах спільної компресії, які відрізняються балансом між обчислювальними витратами та точністю; при цьому встановлено принципову невідповідність класичних методів кластеризації вимогам задач такого типу та обґрунтовано доцільність спеціалізованих алгоритмів, орієнтованих на контроль похибки.
- Проведено системний аналіз сучасних підходів до прунінгу великих мовних моделей і виявлено залежність ефективності компресії від мовних характеристик даних.
- Запропоновано спосіб інтерпретації мов у вигляді геометричних об’єктів, що ґрунтується на структурному аналізі параметрів нейронних мереж, та показано, що відповідний метричний простір відображає змістовні взаємозв’язки між мовами.

- Одержано оцінки впливу процедур прунінгу на якість функціонування нейромережевих контролерів у задачах керування та навчання з підкріпленням, що розширює область застосування отриманих результатів.

Структура дисертаційної роботи вирізняється внутрішньою узгодженістю та чітко простежуваною логікою розвитку результатів: виклад послідовно переходить від систематизації сучасного стану досліджень до отримання нових теоретичних тверджень і їх експериментального підтвердження. Матеріал дисертаційного дослідження викладено англійською мовою. *Перший розділ* має характер розширеного аналітичного огляду, у якому розглянуто як базові принципи побудови нейронних мереж і сучасні архітектури, так і питання оптимального керування та навчання з підкріпленням. Значну увагу приділено методам компресії, зокрема підходам, заснованим на факторизації матриць і прунінгу. Виклад доповнено фундаментальними результатами теорії збурень і низькорангових апроксимацій, які формують математичне підґрунтя подальших досліджень. У цьому ж розділі введено постановку задачі для конкатенованих матриць і розглянуто інкрементальні варіанти сингулярного розкладу, що мають методологічне значення для наступних розділів.

У *другому розділі* розвинуто теоретичний апарат для аналізу похибок, що виникають у задачах зменшення розмірності. Основну увагу приділено дослідженню збурень сингулярного спектра та оцінкам похибки для конкатенованих матриць. Отримані результати, зокрема у теоремі 2.5, встановлюють нетривіальні оцінки, які враховують блокову структуру матриці та виражаються через операторні норми її компонент. На цій основі запропоновано алгоритмічні процедури кластеризації, побудовані за жадібним принципом із використанням критерію допустимої похибки. Подальші теоретичні твердження, зокрема теорема 2.11 та відповідні наслідки, забезпечують оцінки похибки низькорангового наближення, придатні як для теоретичного аналізу, так і для практичної реалізації. Сукупність отриманих результатів дозволяє сформулювати узгоджений критерій об’єднання матриць у кластер із контролем якості, що принципово відрізняє запропонований підхід від класичних методів кластеризації, не орієнтованих на обмеження похибки.

У *третьому розділі* дослідження зосереджено на задачах прунінгу нейронних мереж і їх практичному застосуванні. На основі числових експериментів для великих мовних моделей встановлено, що ефективність прунінгу залежить від мовної специфіки даних, використаних на етапі калібрування. Це спостереження інтерпретовано у вигляді побудови метричного простору мов, визначеного через сигнали важливості параметрів. Подальший теоретичний аналіз спрямовано на встановлення оцінок впливу прунінгу на якість

функціонування моделей у задачах керування. Зокрема, отримано оцінки погіршення якості стратегій керування з урахуванням структурних змін параметрів, а також встановлено узгодження цих результатів із класичними співвідношеннями теорії навчання з підкріпленням. Важливо, що отримані оцінки мають конструктивний характер і можуть бути використані без суттєвого ускладнення обчислювальних процедур.

У сукупності отримані результати є узгодженими між собою та визначають єдину концептуальну основу дослідження компресії моделей машинного навчання. ***Практична цінність*** роботи полягає у можливості застосування цих результатів для розроблення ефективних алгоритмів із гарантованим контролем похибки, що зумовлює їх релевантність для широкого кола прикладних задач, зокрема в обробці природної мови та системах керування складними динамічними об'єктами.

Матеріали дисертації пройшли необхідну ***апробацію*** (5 конференцій, воркшопів, семінарів). Високий рівень виконання наукових досліджень підтверджується якістю публікацій (8 робіт, серед яких 3 проіндексовано у Scopus).

Дисертацію оформлено відповідно до вимог, що висуваються до кваліфікаційних робіт на здобуття ступеня доктора філософії. Порушень академічної доброчесності у дисертації та наукових працях М.Б. Шамрая не виявлено.

Попри те, що дисертаційне дослідження загалом виконано на достатньо високому науковому рівні до дисертаційної роботи є наступні ***зауваження та побажання:***

1. На стор.62 зазначено, що матриця S_c , яка складена з власних значень, є діагональною. Виникає питання, чи буде впливати в подальшому, й якщо так, то як випадок наявності кратних та комплексних власних значень.
2. При доведенні Теорем 2.11, 2.14 враховується ермітовість відповідних матриць. Що робити у випадку наявності комплексних власних значень, якщо такі можуть бути.
3. У підрозділі 2.2.4 проведено ряд числових експериментів на датасетах з найбільшим розміром у 3,83GB. Зважаючи на суттєве зростання обсягів інформації, що сьогодні вимагає обробки, бажано було б провести експерименти для даних, що підпадають під поняття Big Data, й на порядки мають більші об'єми.
4. У розділі 3 наведено Твердження 3.9 про «нерозширюваність» активаційних функцій типу ReLU. Але ж кожна з них має свої специфічні відмінності, бажано було б провести їх порівняльний аналіз.
5. В тексті роботи присутні певні технічні одруківки (наприклад, на стор.87 помилково вказано «Table ??»).

Зазначені зауваження не мають принципового значення та не впливають на цінність та позитивну оцінку дисертації Шамрая М.Б.

На основі наведеного вважаю, що дисертаційна робота «Low-rank approximations, perturbation bounds, and their role in model compression» («Низькорангові наближення, оцінки збурень та їх роль у компресії моделей») і наукові публікації Максима Борисовича Шамрая відповідають спеціальності 113 Прикладна математика та задовольняють вимоги постанови № 44 Кабінету Міністрів України від 12 січня 2022 р. «Про затвердження Порядку присудження ступеня доктора філософії та скасування рішення разової спеціалізованої вченої ради закладу вищої освіти, наукової установи про присудження ступеня доктора філософії», а їх автор заслуговує на присудження ступеня доктора філософії за спеціальністю 113 Прикладна математика в галузі знань 11 Математика та статистика.

Офіційний опонент:
доцент кафедри моделювання складних систем
факультету комп'ютерних наук та кібернетики
Київського національного університету
імені Тараса Шевченка,
доктор фізико-математичних наук, професор
Шатирко Андрій Володимирович

Підпис засвідчую
Вчений секретар НДЧ
Караульна Н.В.
30.04.2026 р.