КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 2 (34), 2026
ISSN 2663 – 4023
DOI 10.28925/2663-4023.2026.34.1360
УДК 004

Богаченко Сергій Володимирович
аспірант кафедри кібербезпеки та програмного забезпечення
Центральноукраїнський національний технічний університет, Кропивницький, Україна
ORCID: 0009-0007-5787-0559
sergeybohachenkodevs@gmail.com

Смірнов Олексій Анатолійович
д.т.н., професор, завідувач кафедри кібербезпеки та програмного забезпечення
Центральноукраїнський національний технічний університет, Кропивницький, Україна
ORCID: 0000-0001-9543-874X
dr.smirnovoa@gmail.com

Буравченко Костянтин Олегович
к.т.н., доцент, доцент кафедри кібербезпеки та програмного забезпечення
Центральноукраїнський національний технічний університет, Кропивницький, Україна
ORCID: 0000-0001-6195-7533
buravchenkok@gmail.com

Смірнова Тетяна Віталіївна
к.т.н., доцент кафедри кібербезпеки та програмного забезпечення
Центрально український національний технічний університет, Кропивницький, Україна
ORCID: 0000-0001-6896-0612
sm.tetyana@gmail.com

Конопліцька-Слободенюк Оксана Костянтинівна
викладач кафедри кібербезпеки та програмного забезпечення
Центральноукраїнський національний технічний університет, Кропивницький, Україна
ORCID: 0000-0001-9981-5194
ksuha80@gmail.com

Якименко Наталія Миколаївна
кандидат фізико-математичних наук, доцент, доцент кафедри кібербезпеки та програмного забезпечення
Центральноукраїнський національний технічний університет, Кропивницький, Україна
ORCID: 0000-0002-4498-0093
yakimenko_n_m@ukr.net

Смірнов Сергій Анатолійович
кандидат технічних наук, доцент, доцент кафедри кібербезпеки та програмного забезпечення
Центральноукраїнський національний технічний університет, Кропивницький, Україна
ORCID: 0000-0002-7649-7442
smirnov.ser.81@gmail.com

ОБМЕЖЕННЯ РЕСУРСІВ МОБІЛЬНИХ ТА EDGE-ПЛАТФОРМ ПРИ РОЗРОБЦІ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ ДЛЯ РЕАЛІЗАЦІЇ ТА ОПТИМІЗАЦІЇ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ

Анотація. У даній роботі проведено аналіз сучасних підходів до реалізації та оптимізації моделей машинного навчання для мобільних та edge-платформ. Актуальність цієї проблематики обумовлена стрімким розвитком мобільних обчислень, систем інтернету речей, периферійних обчислень та необхідністю виконання інтелектуальної обробки даних безпосередньо на пристроях із обмеженими ресурсами. Показано, що сучасні мобільні та edge-пристрої поступово перетворюються на повноцінні обчислювальні системи, які здатні виконувати складні задачі машинного навчання та штучного інтелекту без використання віддалених серверів. Разом із цим використання моделей безпосередньо на мобільних пристроях супроводжується низкою суттєвих обмежень, пов’язаних із обмеженою обчислювальною потужністю, невеликим обсягом оперативної пам’яті, енергетичними

обмеженнями та особливостями теплового режиму роботи. Окремо встановлено, що мобільне середовище характеризується високою динамічністю доступних ресурсів. Доступна обчислювальна потужність, рівень заряду батареї, температура пристрою, активність фонових процесів та доступна пам'ять можуть змінюватися протягом роботи системи. Це створює додаткові труднощі для виконання моделей машинного навчання та суттєво ускладнює використання статичних методів оптимізації.

Ключові слова: обмеження ресурсів, штучний інтелект, машинне навчання, розробка програмного забезпечення, мобільні платформи, edge-платформи.

ВСТУП

Постановка завдання дослідження

Сучасні мобільні та edge-платформи поступово перетворилися з простих користувацьких пристроїв на повноцінні обчислювальні системи, здатні виконувати складні задачі обробки даних, комп'ютерного зору, розпізнавання мовлення та інших напрямів машинного навчання. На відміну від класичних хмарних або серверних рішень, такі пристрої дозволяють виконувати інференс безпосередньо біля джерела даних, що зменшує затримки, підвищує автономність системи та покращує конфіденційність.

Постановка проблеми. Протягом останніх років розвиток мобільних технологій відбувається досить швидкими темпами. Сучасні смартфони, планшети та різноманітні edge-пристрої вже давно перестали використовуватися лише для базових задач зв'язку чи перегляду інформації. Сьогодні вони активно застосовуються у системах комп'ютерного зору, доповненої реальності, біометричної ідентифікації, інтелектуальних системах відеоспостереження та багатьох інших сферах. Досить часто саме мобільний пристрій стає першою ланкою у процесі аналізу та обробки інформації.

Окрему роль у цьому процесі відіграє концепція edge computing. Основна ідея такого підходу полягає у перенесенні обчислень ближче до місця виникнення даних. Це дозволяє зменшити навантаження на мережеву інфраструктуру, скоротити час передачі інформації та забезпечити швидшу реакцію системи. Для задач реального часу це є особливо важливим, оскільки навіть незначні затримки можуть впливати на кінцевий результат роботи системи.

Наприклад, у системах розпізнавання облич, автономної навігації або аналізу відеопотоку необхідно виконувати інференс практично миттєво. Передача всіх даних у хмару може створювати додаткові затримки, а інколи бути взагалі неможливою через нестабільність мережі або відсутність підключення. Саме тому локальне виконання моделей машинного навчання на мобільних пристроях стає дедалі актуальнішим.

У даному дослідженні необхідно розв'язати наступні завдання:

1. Визначити обмеження пам'яті та обчислювальних ресурсів.
2. Розглянути динамічність ресурсів мобільних систем.
3. Провести моніторинг системних ресурсів.

Аналіз останніх досліджень і публікацій. У роботі [1], визначено поняття Edge AI. Edge AI – це використання штучного інтелекту безпосередньо на пристроях (смартфонах, камерах, IoT), без надсилання даних у хмару. Це дозволяє обробляти інформацію та ухвалювати рішення локально й миттєво. Робота [2] визначає, що Edge computing переносить обчислення безпосередньо до джерела даних. Це мінімізує затримки та забезпечує миттєву обробку, що критично для сервісів реального часу, зокрема IoT. У роботі [3] розглянутий принцип роботи Mobile Edge Computing (MEC). У роботі [4] наведені найкращі практики для підвищення продуктивності. TLightweight Network Architecture: оптимізація моделей глибокого навчання розглянута в роботах [5, 6]. У роботі [7] розглянуті моделі з відкритим кодом для ефективних застосувань візуального зору на пристроях. В роботах [8], [9] розглянуто MobileNetV2. У роботі [10] розглянуто EfficientNet: масштабування моделі для згорткових нейронних мереж.

Метою статті є розгляд обмеження ресурсів мобільних та edge-платформ при розробці програмного забезпечення для реалізації та оптимізації моделей машинного навчання та визначення актуальних напрямків подальших досліджень

Об'єктом дослідження є процес машинного навчання.

Предметом дослідження є обмеження ресурсів мобільних та edge-платформ при розробці програмного забезпечення для реалізації та оптимізації моделей машинного навчання.

ТЕОРЕТИЧНІ ОСНОВИ ДОСЛІДЖЕННЯ

Мобільні та edge-пристрої функціонують в умовах значних ресурсних обмежень. До основних таких обмежень належать обмежена обчислювальна потужність, невеликий обсяг доступної оперативної пам'яті, обмежена ємність акумуляторної батареї, теплові обмеження та нестабільність доступних ресурсів у

реальному часі. Саме ці фактори суттєво ускладнюють використання складних моделей глибокого навчання на мобільних платформах.

Продуктивність мобільних систем не можна розглядати як постійну величину. Доступність ресурсів залежить від рівня заряду батареї, температури, фонових процесів та політик операційної системи.

Крім того, мобільні операційні системи самостійно керують значною частиною системних ресурсів. Android та iOS активно використовують механізми динамічного розподілу навантаження, керування пам'яттю та контролю фонової активності. У деяких випадках операційна система може зменшувати доступність окремих апаратних ресурсів для економії енергії або зниження температури.

Важливим обмеженням є обсяг оперативної пам'яті. Для моделей глибокого навчання пам'ять потрібна не лише для зберігання ваг, але й для проміжних активацій, вхідних даних, буферів та службових структур фреймворку. У мобільних системах значна частина пам'яті вже використовується операційною системою та фоновими процесами, тому реальний доступний обсяг пам'яті для моделі може бути значно меншим, ніж загальний обсяг RAM пристрою.

У деяких випадках проміжні результати можуть займати навіть більше пам'яті, ніж самі параметри нейронної мережі. Це особливо характерно для задач комп'ютерного зору, де використовуються великі вхідні зображення або відеопотоки високої роздільної здатності. Через це навіть моделі, які мають відносно невеликий розмір, можуть створювати проблеми при виконанні на пристроях із обмеженим обсягом оперативної пам'яті.

Окрему роль відіграє енергоспоживання. На серверних платформах енергетичні витрати часто оцінюються з погляду загальної ефективності дата-центру, тоді як для мобільного пристрою енергоспоживання безпосередньо впливає на час автономної роботи. Інтенсивний інференс може швидко розряджати батарею, викликати нагрівання пристрою та погіршувати користувацький досвід.

Проблема енергоспоживання стає ще помітнішою у випадках безперервного виконання інференсу. Наприклад, системи аналізу відеопотоку, детектування об'єктів або розпізнавання жестів можуть виконувати обчислення практично постійно. За таких умов навіть незначне збільшення складності моделі здатне суттєво впливати на загальний час автономної роботи пристрою.

Тепловий фактор також є суттєвим. Через компактність корпусу мобільні пристрої мають обмежені можливості відведення тепла. У разі тривалого навантаження операційна система може знижувати частоту CPU або GPU, щоб запобігти перегріву. Це явище відоме як thermal throttling. Для задач машинного навчання воно є особливо критичним, оскільки може призводити до різкого збільшення часу інференсу.

Особливість теплового тротлінгу полягає у тому, що продуктивність системи може змінюватися вже під час роботи алгоритму. Наприклад, модель може демонструвати хорошу швидкість виконання протягом перших хвилин, але після нагрівання процесора її продуктивність починає поступово знижуватися. Через це результати тестування моделей у лабораторних умовах не завжди відповідають реальним сценаріям використання.

У контексті edge computing проблема стає ще ширшою. Обчислення можуть виконуватися не лише на мобільному пристрої, а й на edge-сервері або у хмарі. Такий підхід дозволяє розподіляти навантаження, але створює нові виклики: потрібно враховувати затримку мережі, стабільність з'єднання, енергетичні витрати на передачу даних та вимоги до конфіденційності.

При цьому рішення щодо того, де саме повинна виконуватися модель, не завжди є очевидним. У деяких випадках локальне виконання може бути більш енергоефективним, а в інших навпаки доцільно передати частину обчислень на зовнішній ресурс. Це додатково ускладнює організацію роботи систем машинного навчання.

Отже, мобільні та edge-платформи формують складне середовище для виконання моделей машинного навчання. У такому середовищі важливо не лише зменшити розмір моделі, але й забезпечити її здатність працювати стабільно в умовах змінних ресурсів. Саме тому сучасні дослідження дедалі більше орієнтуються не лише на створення компактних архітектур, а й на розробку адаптивних методів виконання інференсу.

Отже, мобільні та edge-платформи мають ряд особливостей, які суттєво відрізняють їх від класичних серверних систем. Обмеженість пам'яті, нестабільність ресурсів, енергетичні та теплові фактори формують складне середовище виконання моделей машинного навчання. Це створює необхідність розробки нових підходів до оптимізації та адаптивного управління інференсом в умовах обмежених ресурсів.

Обмеження пам'яті та обчислювальних ресурсів

Одним із головних факторів, що впливають на ефективність виконання моделей машинного навчання на мобільних та edge-пристроях, є обмеженість пам'яті та обчислювальних ресурсів. Саме ця проблема значною мірою визначає можливість практичного використання сучасних алгоритмів штучного

інтелекту в умовах обмежених ресурсів. Якщо у серверних середовищах нестача пам'яті або недостатня продуктивність часто вирішуються шляхом масштабування інфраструктури, то у мобільних системах така можливість суттєво обмежена або взагалі відсутня.

Моделі глибокого навчання, особливо згорткові нейронні мережі та сучасні трансформерні архітектури, можуть містити мільйони або навіть десятки мільйонів параметрів. Зі збільшенням складності моделі зростають вимоги як до пам'яті, так і до обчислювальної потужності. При цьому зростання кількості параметрів не завжди відбувається лінійно відносно приросту точності, через що використання великих моделей на мобільних пристроях часто стає неефективним.

Однак самі параметри моделі є лише частиною загального споживання ресурсів. Крім ваг нейронної мережі, значний обсяг пам'яті використовується для зберігання проміжних результатів обчислень, активацій, буферів, службових структур та внутрішніх компонентів фреймворків машинного навчання. У деяких випадках саме ці проміжні дані створюють основне навантаження на пам'ять.

У процесі інференсу кожен шар нейронної мережі формує набір активацій, які можуть займати більше пам'яті, ніж самі ваги моделі. Це особливо помітно у задачах комп'ютерного зору, де вхідні дані мають високу роздільну здатність. Наприклад, при роботі з відеопотоками або зображеннями великого розміру кількість проміжних тензорів може стрімко збільшуватися. Через це зменшення кількості параметрів моделі не завжди автоматично означає пропорційне зменшення використання пам'яті.

Окрему проблему створює той факт, що частина оперативної пам'яті мобільного пристрою вже використовується системними службами та операційною системою. Android та iOS виконують значну кількість процесів у фоновому режимі: синхронізацію даних, мережеві сервіси, роботу інтерфейсу, геолокаційні служби та інші системні задачі. Через це реальний доступний обсяг пам'яті може бути значно меншим за формально заявлений виробником пристрою.

Наприклад, навіть якщо смартфон оснащений 8 ГБ оперативної пам'яті, доступний обсяг для застосунку може бути суттєво нижчим. Частина пам'яті резервується під графічну систему, фонові процеси та внутрішні механізми керування ресурсами. Для моделей машинного навчання це створює додаткові труднощі, особливо у випадках, коли одночасно виконуються інші ресурсоємні задачі.

Обчислювальні ресурси мобільних пристроїв також мають низку особливостей. Сучасні смартфони часто використовують гетерогенні процесорні архітектури, де поєднуються продуктивні та енергоефективні ядра. Такий підхід дозволяє балансувати між швидкодією та енергоспоживанням, але водночас ускладнює прогнозування продуктивності моделі.

У більшості сучасних процесорів використовується підхід big.LITTLE, відповідно до якого частина ядер орієнтована на максимальну продуктивність, а інша частина оптимізована для енергоощадної роботи. Операційна система самостійно визначає, на яких ядрах будуть виконуватися задачі залежно від поточного навантаження та режиму роботи пристрою. Через це одна і та сама операція може виконуватися з різною швидкістю залежно від того, на якому типі ядра вона була запущена. У реальних умовах це може створювати нестабільність часу інференсу, особливо при тривалому виконанні алгоритмів машинного навчання.

Крім CPU, сучасні мобільні платформи можуть містити GPU та NPU. Використання таких прискорювачів дозволяє суттєво підвищити продуктивність виконання нейронних мереж, оскільки значна частина операцій машинного навчання добре підходить для паралельної обробки.

GPU традиційно ефективно працюють із матричними обчисленнями та великими обсягами однотипних операцій. У свою чергу NPU є спеціалізованими апаратними модулями, оптимізованими безпосередньо під задачі нейронних мереж. У деяких випадках використання NPU дозволяє суттєво зменшити енергоспоживання та скоротити час виконання інференсу.

Разом із цим ефективність таких прискорювачів суттєво залежить від підтримки конкретних операцій, фреймворків та оптимізованих бібліотек. Якщо модель містить операції, які не підтримуються апаратним прискорювачем, частина обчислень може виконуватися на CPU, що призводить до погіршення загальної продуктивності.

У деяких випадках часткове використання GPU або NPU може навіть створювати додаткові накладні витрати, пов'язані з передачею даних між обчислювальними блоками. Через це не кожна модель автоматично отримує переваги від використання спеціалізованих прискорювачів.

Важливо зазначити, що обмеження ресурсів мають не лише кількісний, але й якісний характер. Наприклад, модель може мати невелику кількість параметрів, але через неефективний доступ до пам'яті або слабку підтримку операцій на конкретному пристрої працювати повільніше, ніж більша, але краще оптимізована модель.

Саме тому при аналізі ефективності моделей для мобільних платформ недостатньо враховувати лише кількість параметрів або FLOPs. Такі метрики є важливими, однак вони не відображають реальну поведінку системи в умовах практичного використання.

З огляду на це, для мобільних та edge-систем важливо оцінювати модель комплексно: за розміром, часом інференсу, споживанням пам’яті, енергоспоживанням та стабільністю роботи в реальних умовах. Лише комплексний підхід дозволяє об’єктивно оцінити придатність моделі до використання на пристроях із обмеженими ресурсами.

Таким чином, обмеження пам’яті та обчислювальних ресурсів є одними з основних факторів, що впливають на ефективність виконання моделей машинного навчання на мобільних та edge-платформах. Важливо враховувати не лише кількість параметрів або теоретичну складність моделі, але і особливості апаратної архітектури, реальне використання пам’яті та підтримку апаратних прискорювачів. Це створює потребу у використанні комплексних підходів до оптимізації моделей та адаптивного керування процесом інференсу.

Динамічність ресурсів мобільних систем

У більшості сучасних мобільних платформ доступні обчислювальні ресурси не можна вважати статичними. Вони можуть змінюватися не лише під впливом користувача, але й через внутрішні механізми операційної системи, поточний стан апаратних компонентів або особливості мережевого середовища. Через це одна і та сама модель машинного навчання може демонструвати різні характеристики продуктивності навіть на одному пристрої та за однакових вхідних даних.

Однією з особливостей мобільного середовища є те, що зміни стану ресурсів можуть відбуватися досить швидко. Якщо на серверних системах зміна навантаження часто прогнозована, то у мобільних системах вона може виникати раптово. Наприклад, користувач може отримати відеодзвінок, увімкнути GPS-навігацію, розпочати запис відео або одночасно відкрити декілька застосунків. У таких умовах доступний обсяг ресурсів для моделей машинного навчання може змінитися буквально за декілька секунд.

Наприклад, якщо користувач одночасно використовує камеру, навігацію, мобільний інтернет і застосунок з ML-моделлю, доступні ресурси для інференсу можуть суттєво зменшуватися. Це пов'язано з тим, що кожен із таких сервісів використовує процесорний час, пам’ять, графічні ресурси та енергетичні можливості пристрою. Внаслідок цього модель машинного навчання змушена працювати вже не в оптимальних умовах.

Досить важливим фактором є робота операційної системи. Android та iOS самостійно керують процесами, пам’яттю, пріоритетами задач та енергоспоживанням. Сучасні мобільні системи використовують складні механізми керування ресурсами, які постійно аналізують стан пристрою та приймають рішення щодо розподілу доступних обчислювальних можливостей.

У разі високого навантаження або низького заряду батареї система може обмежувати фонову активність, знижувати частоту процесора або змінювати доступність апаратних прискорювачів. Наприклад, під час переходу пристрою в режим енергозбереження деякі високопродуктивні ядра можуть використовуватися рідше або працювати на нижчих частотах. Подібні механізми дозволяють збільшити час автономної роботи, але водночас можуть негативно впливати на швидкодію інференсу.

Окремо необхідно враховувати температурний стан пристрою. У процесі виконання складних обчислень процесор, GPU та інші компоненти можуть нагріватися. При досягненні певного температурного порогу система автоматично активує механізми обмеження продуктивності. У результаті частота процесора знижується, що призводить до зміни швидкодії моделі.

Особливість полягає у тому, що такі зміни можуть відбуватися поступово та бути майже непомітними для користувача. Однак для систем машинного навчання навіть незначне зниження продуктивності може впливати на затримку інференсу, особливо у випадку задач реального часу.

Ще одним прикладом динамічності є мережеве середовище. У задачах, де використовується edge-cloud інференс, якість мережевого з’єднання має прямий вплив на продуктивність системи. Перемикання між Wi-Fi, 4G або 5G, а також нестабільний сигнал можуть збільшувати затримку та енергетичні витрати на передачу даних.

У сучасних системах мобільного машинного навчання частина обчислень може виконуватися локально, а інша частина передаватися на edge-сервер або у хмарну інфраструктуру. При цьому якість мережевого з’єднання стає важливим фактором, який впливає на ефективність такого підходу. Якщо канал зв’язку нестабільний або має високу затримку, переваги від розподіленого виконання можуть суттєво зменшуватися.

Також слід враховувати, що передача великих обсягів даних через мобільні мережі сама по собі створює додаткове енергетичне навантаження. Через це інколи локальний інференс може виявитися більш ефективним навіть при меншій продуктивності пристрою.

Слід зазначити, динамічність ресурсів є не другорядною проблемою, а базовою властивістю мобільного середовища. Саме вона створює потребу в адаптивних методах виконання моделей, які можуть змінювати свою поведінку залежно від поточного стану пристрою.

Традиційні статичні методи оптимізації зазвичай орієнтуються на фіксовані характеристики обладнання. Однак для мобільного середовища такий підхід часто виявляється недостатнім. Саме тому сучасні дослідження дедалі частіше спрямовані на розробку адаптивних алгоритмів, які здатні змінювати параметри виконання моделі відповідно до поточних умов роботи системи.

Варто зазначити, динамічність ресурсів є однією з базових характеристик мобільних систем, яка суттєво ускладнює виконання моделей машинного навчання. Нестабільність обчислювальних ресурсів, зміни температурного режиму, рівня заряду батареї та мережевих умов створюють середовище, у якому статичні підходи до виконання моделей не завжди забезпечують необхідну ефективність. Це формує потребу у використанні адаптивних стратегій інференсу, здатних враховувати поточний стан системи.

Моніторинг системних ресурсів

Моніторинг системних ресурсів є необхідною основою для побудови адаптивних методів машинного навчання на мобільних та edge-платформах. Без інформації про поточний стан пристрою неможливо приймати обґрунтовані рішення щодо зміни параметрів інференсу, вибору апаратного прискорювача або зменшення обчислювальної складності моделі. Фактично моніторинг виступає механізмом, який дозволяє системі отримувати інформацію про власний стан та реагувати на зміни навколишнього середовища.

У традиційних серверних системах обчислювальні ресурси часто мають відносно стабільний характер, тому потреба у постійному динамічному аналізі стану обладнання не завжди є критичною. Для мобільних платформ ситуація суттєво відрізняється. Як уже зазначалося раніше, доступні ресурси можуть змінюватися в режимі реального часу залежно від навантаження, температури, рівня заряду батареї або дій користувача. Через це система повинна постійно отримувати актуальні дані про свій стан.

Без механізмів моніторингу реалізація адаптивного інференсу стає практично неможливою. Якщо система не знає, наскільки завантажений процесор або скільки пам’яті залишилося доступною, вона не може приймати рішення щодо зміни конфігурації моделі або способу виконання обчислень. У результаті навіть оптимізована модель може працювати неефективно.

До основних параметрів, які доцільно відстежувати, належать завантаження CPU, GPU та NPU, доступний обсяг оперативної пам’яті, температура пристрою, рівень заряду батареї, поточне енергоспоживання та характеристики мережевого з’єднання. У реальних системах частина цих показників може бути доступною через API операційної системи, а частина потребує профілювання або непрямої оцінки.

Слід зазначити, що не всі характеристики можуть бути отримані безпосередньо. Наприклад, деякі виробники мобільних платформ обмежують доступ до внутрішніх показників енергоспоживання або температури окремих компонентів. Через це у практичних системах часто використовуються наближені методи оцінювання або непрямі показники.

Моніторинг CPU дозволяє оцінити поточне навантаження на процесорні ядра та визначити, чи може система виконувати додаткові обчислення без ризику затримок. Особливо це важливо для задач реального часу, де навіть незначне перевантаження процесора може впливати на швидкодію системи.

Крім загального завантаження процесора, додатково може аналізуватися використання окремих ядер, зміна частоти процесора або розподіл навантаження між продуктивними та енергоефективними ядрами. Це важливо через те, що сучасні мобільні процесори використовують гетерогенні архітектури, а продуктивність окремих ядер може суттєво відрізнятися.

Моніторинг пам’яті дає змогу уникати ситуацій, коли модель або застосунок починають конкурувати з іншими процесами за RAM. Нестача пам’яті може призводити до сповільнення роботи системи, додаткових звернень до механізмів очищення пам’яті або навіть примусового завершення процесів операційною системою.

У випадку великих моделей або задач комп’ютерного зору ця проблема стає особливо помітною. Оперативна пам’ять використовується не лише для ваг нейронної мережі, але і для великої кількості проміжних активацій, тензорів та буферів. Через це навіть короткочасні зміни доступного обсягу пам’яті можуть впливати на стабільність роботи інференсу.

Контроль температури важливий для запобігання тротлінгу, а відстеження стану батареї дозволяє переходити до енергоощадних режимів роботи. Якщо система виявляє швидке зростання температури або низький рівень заряду, вона може заздалегідь змінити параметри виконання моделі, не очікуючи моменту, коли операційна система примусово знизить продуктивність пристрою.

Окрему увагу необхідно приділяти мережевим характеристикам. У випадку використання edge-cloud інференсу швидкість передачі даних, затримка мережі та стабільність з’єднання безпосередньо впливають на ефективність виконання задачі. Якщо затримка мережі перевищує певне значення, передача даних на віддалений сервер може виявитися менш ефективною, ніж локальний інференс.

У контексті машинного навчання моніторинг має не лише діагностичну, але й керуючу роль. Зібрані метрики можуть використовуватися для вибору конфігурації моделі, частоти інференсу, точності обчислень або апаратного ресурсу.

Наприклад, при високій температурі пристрою система може перейти на легшу модель або зменшити частоту обробки кадрів. У разі низького заряду батареї можуть використовуватися енергоощадні режими обчислень або знижена точність операцій. Аналогічно при високому навантаженні CPU інференс може бути перенаправлений на GPU або NPU. Моніторинг створює механізм зворотного зв'язку між станом системи та поведінкою моделі.

Отже, моніторинг ресурсів є проміжною ланкою між станом мобільного пристрою та механізмами адаптації моделі. Саме ця ланка дозволяє перейти від статичного інференсу до динамічного, ресурсно-орієнтованого виконання.

Підсумовуючи, моніторинг системних ресурсів є необхідною складовою сучасних систем мобільного машинного навчання. Його використання дозволяє враховувати поточний стан пристрою та адаптувати виконання моделей відповідно до змін умов роботи. Це створює основу для реалізації динамічного та ресурсно-орієнтованого інференсу, який здатний забезпечити баланс між продуктивністю, точністю та енергоспоживанням.

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

У даній роботі досліджено сучасні підходи до побудови та оптимізації моделей машинного навчання для мобільних і периферійних (edge) платформ. Актуальність теми зумовлена масовим впровадженням систем інтернету речей та необхідністю локальної обробки даних на ресурсно обмежених пристроях. Визначено, що попри зростання обчислювальної спроможності edge-пристроїв до рівня автономних інтелектуальних систем, їх ефективне використання стримується лімітами оперативної пам'яті, енергоспоживання та тепловиділення. Окрему увагу приділено динамічності мобільного середовища, де коливання обсягу вільних ресурсів, рівня заряду акумулятора й фонового навантаження ускладнюють застосування статичних методів оптимізації.

У результаті проведеного дослідження визначено перспективний напрям подальших досліджень, який полягає у розробленні адаптивного ресурсно-орієнтованого підходу до керування процесом інференсу моделей машинного навчання.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. InviAI. (n.d.). *Що таке Edge AI? Переваги та виклики*. https://inviai.com/uk/sho-take-edge-ai
2. ProIT. (n.d.). *Як edge computing змінює обчислювальні технології*. https://proit.com.ua/news/yak-edge-computing-zminyuye-obchyslyuvalni/
3. ETSI. (2019). *Mobile Edge Computing (MEC); Framework and reference architecture* (ETSI GS MEC 003 V2.1.1). https://www.etsi.org/deliver/etsi_gs/MEC/001_099/003/02.01.01_60/gs_MEC003v020101p.pdf
4. Google. (n.d.). *TensorFlow Lite guide: Performance best practices*. https://www.tensorflow.org/lite/performance/best_practices
5. Mao, Z. (2025). Lightweight network architecture – Optimization of deep learning models. *Applied and Computational Engineering*. https://doi.org/10.54254/2755-2721/2025.AST26509
6. Liu, L., & Xu, Z. (2025). Optimizing lightweight neural networks for efficient mobile edge computing. *Scientific Reports, 15*, 22056. https://doi.org/10.1038/s41598-025-04652-7
7. Howard, A., Zhu, M., Chen, B., et al. (2017). *MobileNets: Open-source models for efficient on-device vision applications*. Google Research. https://research.google/blog/mobilenets-open-source-models-for-efficient-on-device-vision/
8. Sandler, M., Howard, A., Zhu, M., et al. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *IEEE/CVF CVPR*. https://doi.org/10.48550/arXiv.1801.04381
9. Ma, N., Zhang, X., Zheng, H., & Sun, J. (2018). ShuffleNet V2: Practical guidelines for efficient CNN architecture design. In *ECCV*. https://link.springer.com/chapter/10.1007/978-3-030-01264-9_8
10. Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *ICML*. https://proceedings.mlr.press/v97/tan19a.html
11. Kuznetsov, O., Frontoni, E., Kuznetsova, Y., Chevardin, V., & Smirnov, O. (2025). Architectural foundations for adaptive security in edge computing systems. In *Cybersecurity Defensive Walls in Edge Computing* (pp. 21–61). https://www.scopus.com/pages/publications/105026847051

12. Kuznetsov, O., Atzeni, G., Arnesano, M., Randieri, C., & Smirnov, O. (2025). Secure IoT-based smart wheelchair system: From implementation to security enhancement strategy. In *Security and Privacy of Cyber Physical Systems Emerging Trends Technologies and Applications* (pp. 225–257). https://www.scopus.com/pages/publications/105014300075
13. Kuznetsov, O., Smirnov, O., Kuznetsova, T., Shaikhanova, A., & Svatowsky, I. (2025). Privacy-utility trade-offs in IoT networks: A comparative analysis of differential privacy mechanisms for sensor data aggregation. In *Security and Privacy of Cyber Physical Systems Emerging Trends Technologies and Applications* (pp. 589–622). https://www.scopus.com/pages/publications/105014301897
14. Smirnov, O., Tkachuk, R., Kozirova, N., Konstantynova, L., Konoplitska-Slobodeniuk, O., Yakymenko, N., & Smirnov, S. (2026). Research on the application of vector databases in generative artificial intelligence. *Cybersecurity: Education, Science, Technique, 1*(33), 667–682. https://doi.org/10.28925/2663-4023.2026.33.1248
15. Smirnov, O., Zaritskyi, V., Buravchenko, K., Konoplitska-Slobodeniuk, O., Konstantynova, L., Yakymenko, N., & Smirnov, S. (2026). Optimization of face recognition using the CUDA-accelerated dlib library. *Cybersecurity: Education, Science, Technique, 4*(32), 573–582. https://doi.org/10.28925/2663-4023.2026.32.1154

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 2 (34), 2026
ISSN 2663 – 4023

Serhii Bohachenko
PhD graduate student of the Cybersecurity & Software Academic Department
Central Ukrainian National Technical University, Kropyvnytskyi, Ukraine
ORCID: 0009-0007-5787-0559
sergeybohachenkodevs@gmail.com

Oleksii Smirnov
Doctor of Technical Sciences, Professor,
Head of Cybersecurity & Software Academic Department
Central Ukrainian National Technical University, Kropyvnytskyi, Ukraine
ORCID: 0000-0001-9543-874X
dr.smirnovoa@gmail.com

Kostiantyn Buravchenko
Candidate of Technical Sciences, Associate Professor,
Associate Professor of Cybersecurity & Software Academic Department
Central Ukrainian National Technical University, Kropyvnytskyi, Ukraine
ORCID: 0000-0001-6195-7533
buravchenkok@gmail.com

Tetiana Smirnova
PhD of Science (Engineering),
Associate Professor of Cybersecurity & Software Academic Department
Central Ukrainian National Technical University, Kropyvnytskyi, Ukraine
ORCID:0000-0001-6896-0612
sm.tetyana@gmail.com

Oksana Konoplitska-Slobodeniuk
Lecturer
Central Ukrainian National Technical University, Kropyvnytskyi, Ukraine
ORCID: 0000-0001-9981-5194
ksuha80@gmail.com

Nataliia Yakymenko
candidate of Physical and Mathematical Sciences,
associate professor, associate professor of Cybersecurity & Software Academic Department
Central Ukrainian National Technical University, Kropivnitskiy, Ukraine
ORCID: 0000-0002-4498-0093
yakimenko_n_m@ukr.net

Serhii Smirnov
Candidate of Science (Engineering), associate professor,
associate professor of Cybersecurity & Software Academic Department
Central Ukrainian National Technical University, Kropyvnytskyi, Ukraine
ORCID: 0000-0002-7649-7442
smirnov.ser.81@gmail.com

LIMITATIONS OF RESOURCES OF MOBILE AND EDGE PLATFORMS WHEN DEVELOPING SOFTWARE FOR IMPLEMENTATION AND OPTIMIZATION OF MACHINE LEARNING MODELS

Abstract. This paper analyzes modern approaches to the implementation and optimization of machine learning models for mobile and edge platforms. The relevance of this issue is due to the rapid development of mobile computing, Internet of Things systems, peripheral computing, and the need to perform intelligent data processing directly on devices with limited resources. It is shown that modern mobile and edge devices are gradually transforming into full-fledged computing systems that are capable of performing complex machine learning and artificial intelligence tasks without the use of remote servers. At the same time, the use of models directly on mobile devices is accompanied by a number of significant limitations associated with limited computing power, a small amount of RAM, energy constraints, and the peculiarities of the thermal mode of operation. It is separately established that the mobile environment is characterized by high dynamics of available resources. Available

computing power, battery charge level, device temperature, activity of background processes, and available memory can change during system operation. This creates additional difficulties for the execution of machine learning models and significantly complicates the use of static optimization methods.

Keywords: resource constraints, artificial intelligence, machine learning, software development, mobile platforms, edge platforms.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. InviAI. (n.d.). *What is Edge AI? Benefits and challenges*. https://inviai.com/uk/sho-take-edge-ai
2. ProIT. (n.d.). *How edge computing is changing computing technology*. https://proit.com.ua/news/yak-edge-computing-zminyuye-obchyslyuvalni/
3. ETSI. (2019). *Mobile Edge Computing (MEC); Framework and reference architecture* (ETSI GS MEC 003 V2.1.1). https://www.etsi.org/deliver/etsi_gs/MEC/001_099/003/02.01.01_60/gs_MEC003v020101p.pdf
4. Google. (n.d.). *TensorFlow Lite guide: Performance best practices*. https://www.tensorflow.org/lite/performance/best_practices
5. Mao, Z. (2025). Lightweight network architecture – Optimization of deep learning models. *Applied and Computational Engineering*. https://doi.org/10.54254/2755-2721/2025.AST26509
6. Liu, L., & Xu, Z. (2025). Optimizing lightweight neural networks for efficient mobile edge computing. *Scientific Reports, 15*, 22056. https://doi.org/10.1038/s41598-025-04652-7
7. Howard, A., Zhu, M., Chen, B., et al. (2017). *MobileNets: Open-source models for efficient on-device vision applications*. Google Research. https://research.google/blog/mobilenets-open-source-models-for-efficient-on-device-vision/
8. Sandler, M., Howard, A., Zhu, M., et al. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *IEEE/CVF CVPR*. https://doi.org/10.48550/arXiv.1801.04381
9. Ma, N., Zhang, X., Zheng, H., & Sun, J. (2018). ShuffleNet V2: Practical guidelines for efficient CNN architecture design. In *ECCV*. https://link.springer.com/chapter/10.1007/978-3-030-01264-9_8
10. Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *ICML*. https://proceedings.mlr.press/v97/tan19a.html
11. Kuznetsov, O., Frontoni, E., Kuznetsova, Y., Chevardin, V., & Smirnov, O. (2025). Architectural foundations for adaptive security in edge computing systems. In *Cybersecurity Defensive Walls in Edge Computing* (pp. 21–61). https://www.scopus.com/pages/publications/105026847051
12. Kuznetsov, O., Atzeni, G., Arnesano, M., Randieri, C., & Smirnov, O. (2025). Secure IoT-based smart wheelchair system: From implementation to security enhancement strategy. In *Security and Privacy of Cyber Physical Systems Emerging Trends Technologies and Applications* (pp. 225–257). https://www.scopus.com/pages/publications/105014300075
13. Kuznetsov, O., Smirnov, O., Kuznetsova, T., Shaikhanova, A., & Svatowsky, I. (2025). Privacy-utility trade-offs in IoT networks: A comparative analysis of differential privacy mechanisms for sensor data aggregation. In *Security and Privacy of Cyber Physical Systems Emerging Trends Technologies and Applications* (pp. 589–622). https://www.scopus.com/pages/publications/105014301897
14. Smirnov, O., Tkachuk, R., Kozirova, N., Konstantinova, L., Konoplytska-Slobodeniuk, O., Yakymenko, N., & Smirnov, S. (2026). Research on the application of vector databases in generative artificial intelligence. *Cybersecurity: Education, Science, Technology, 1*(33), 667–682. https://doi.org/10.28925/2663-4023.2026.33.1248
15. Smirnov, O., Zaritsky, V., Buravchenko, K., Konoplytska-Slobodeniuk, O., Konstantinova, L., Yakymenko, N., & Smirnov, S. (2026). Optimization of face recognition using the CUDA accelerated dlib library. *Cybersecurity: Education, Science, Technology, 4*(32), 573–582. https://doi.org/10.28925/2663-4023.2026.32.1154

Отримано редакцією журналу / Received: 17.02.26
Прорецензовано / Revised: 26.02.26
Схвалено до друку / Accepted: 24.09.26