НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»

Факультет інформатики · Кафедра інформатики

ОПТИМІЗАЦІЯ НЕЙРОННОЇ МЕРЕЖІ ДЛЯ РОБОТИ В РОЗШИРЕННІ iOS-ЗАСТОСУНКУ

Кваліфікаційна робота на здобуття освітнього ступеня бакалавр

Виконав: **Гібський Владислав Ігорович**
ІПЗ-4, спеціальність 121

Керівник: **ст. викл. Франків О. О.**

Київ · 2026

Актуальність дослідження

Конфлікт між запитом на приватний on-device AI та обмеженнями iOS-розширень

01

Хмарні рішення

Wispr Flow, Grammarly, SwiftKey — повна залежність від мережі, ризики розкриття приватного контенту, мережеві затримки.

02

Apple Intelligence

Foundation Models доступний лише на iPhone 15 Pro та новіших — виключає значну частку активних пристроїв.

03

Власні локальні моделі

Жодне рішення на ринку не розгортає власну LLM у iOS-розширенні. Технічна неможливість не підтверджена і не спростована.

Виявлений розрив: відсутні рішення, що поєднують приватність, офлайн-роботу та широку сумісність — це і визначає актуальність дослідження.

~1 ГБ

ваги Qwen2.5-0.5B у FP16

77 МБ

ліміт пам'яті клавіатури iOS

>10×

перевищення без оптимізації

Мета і завдання дослідження

МЕТА

Розробити та експериментально перевірити підхід до оптимізації компактних генеративних нейронних мереж для їх стабільного функціонування в умовах ресурсних обмежень iOS-розширень.

Об'єкт: процес розгортання компактних генеративних нейронних мереж у середовищі iOS-розширень.

Предмет: методи оптимізації трансформерних моделей під ліміти phys_footprint розширень.

ЗАВДАННЯ

1. Проаналізувати обмеження iOS-розширень та сучасний інструментарій on-device AI.
2. Систематизувати методи стиснення трансформерних моделей: квантизація, палетизація, прунінг.
3. Порівняти Core ML та MLX у контексті лімітів пам'яті процесів-розширень.
4. Розробити прототип з Custom Keyboard Extension та інтегрованою моделлю.
5. Експериментально виміряти phys_footprint лімітів та ключові метрики виконання.
6. Сформулювати рекомендації щодо розгортання LLM у розширеннях iOS.

Обмеження iOS-розширень

Архітектурний контекст: чому розгортання LLM у розширеннях не тривіальне

АРХІТЕКТУРА ПРОЦЕСІВ

Основний застосунок

Ліміт пам'яті — гігабайти. Повний доступ до GPU та ANE.

IPC / App Group

Розширення (Extension)

- ліміти пам'яті Jetsam: 77–760 МБ
- немає фонового виконання
- обмежений пріоритет до GPU
- ізольована пісочниця

JETSAM

Системний механізм iOS, що примусово завершує процеси при перевищенні встановленого ліміту фізичної пам'яті.

ключова метрика:

`phys_footprint`

Працює на рівні ядра, обробити неможливо. Поточне значення доступне через task_info(TASK_VM_INFO).

Висновок: перевищення phys_footprint навіть на одну сторінку — миттєве завершення процесу.

Інструментарій on-device AI на iOS

Вибір фреймворку напряму визначає, чи зможе модель уписатися в ліміти розширення

Core ML

CPU + GPU + ANE

✓ Сумісний

ПЕРЕВАГИ

+ Виконання на ANE поза процесом
+ Підтримка stateful моделей
+ Інтегрований у систему

ОБМЕЖЕННЯ

− Статичний граф
− Складна конвертація

MLX

тільки GPU (Metal)

✗ Несумісний

ПЕРЕВАГИ

+ Найвища швидкість на GPU
+ Динамічний граф
+ Swift API

ОБМЕЖЕННЯ

− Немає доступу до ANE на iOS
− Ваги повністю у phys_footprint

Foundation Models

ANE

△ Обмежений

ПЕРЕВАГИ

+ Системна модель ~3B
+ Безкоштовно
+ Виконання на ANE

ОБМЕЖЕННЯ

− Тільки iPhone 15 Pro+
− Фіксована системна модель

Методи стиснення нейронних мереж

Що зменшити, як зменшити та який метод оптимальний для виконання на ANE

Квантизація

СТАНДАРТ

Перетворення FP-ваг у цілочислове представлення нижчої розрядності

```
Q(r) = round(r/S) − Z
```

оптимум:

FP32 → INT8 або INT4

Палетизація

ОПТИМУМ ДЛЯ ANE

Кластеризація ваг через k-середніх + таблиця центроїдів

```
2^N центроїдів + індекси
```

оптимум:

4-бітна палетизація

Прунінг

НЕ ДЛЯ ANE

Видалення ваг з найменшим впливом на якість

```
розріджений тензор
```

оптимум:

50% розрідженість

Висновок розділу 2: обрано 4-бітну палетизацію — рекомендована Core ML Tools як найефективніший метод для виконання на ANE.

APPLE NEURAL ENGINE

Ваги моделі виконуються поза процесом застосунку

і тому не входять до phys_footprint — це теоретичний фундамент усієї роботи

ПРОЦЕС ЗАСТОСУНКУ

phys_footprint

Код застосунку + дрібні буфери

KV-кеш та проміжні тензори

✓ Вписується в ліміт Jetsam

shared buffers

СИСТЕМНИЙ ДЕМОН aned

поза процесом

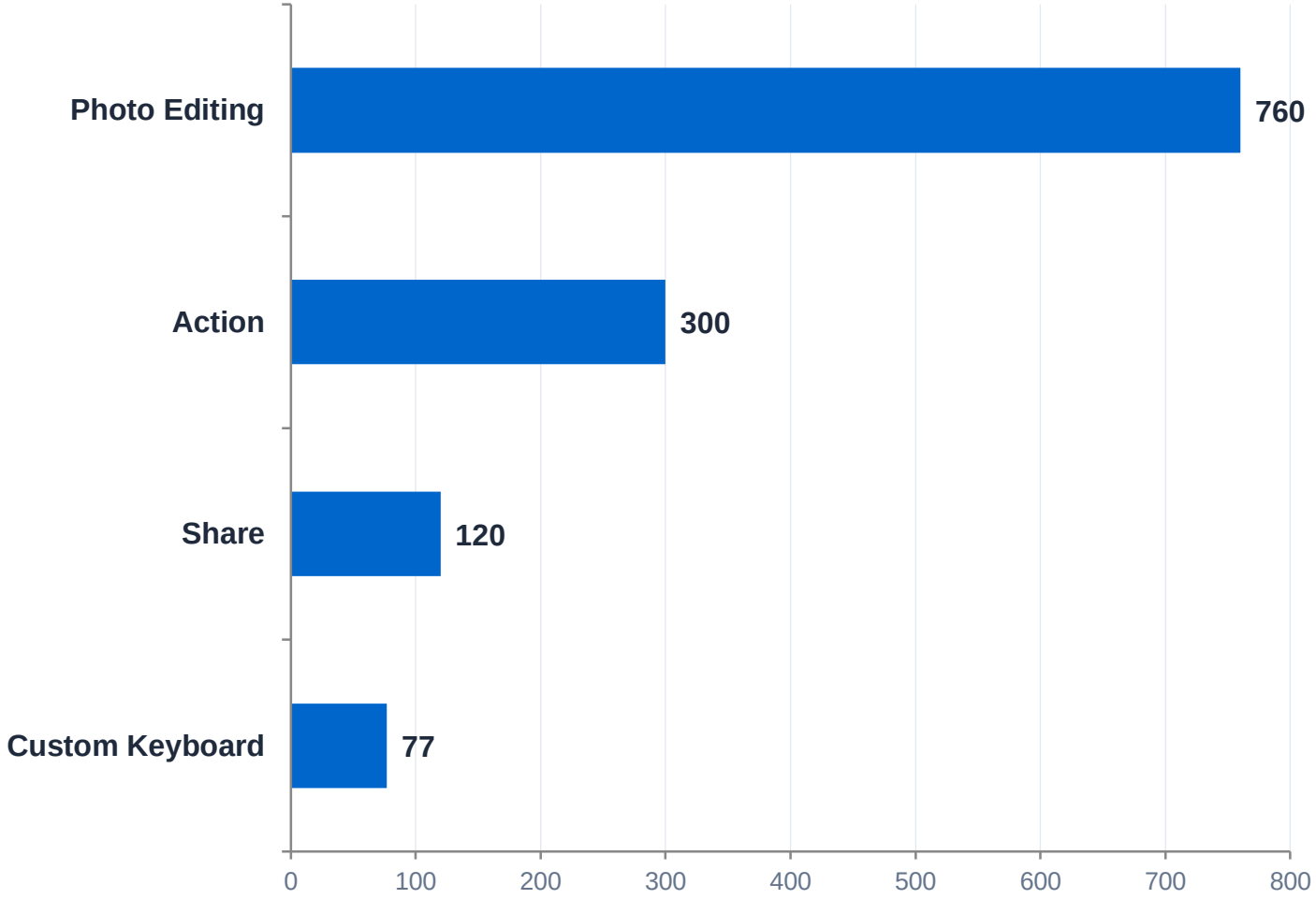
Ваги моделі
(основний обсяг пам'яті)

✓ Не входить у phys_footprint

Передбачення для перевірки: Core ML + cpuAndNeuralEngine → ваги назовні → phys_footprint різко нижчий за розмір моделі.

Експеримент 1: фактичні ліміти розширень

Виміряно через спрацювання Jetsam на iPhone 13 Pro, iOS 26.4. Цикл виділення пам'яті до примусового завершення процесу.



ВИБІР ЦІЛЬОВОГО ТИПУ

Custom Keyboard Extension

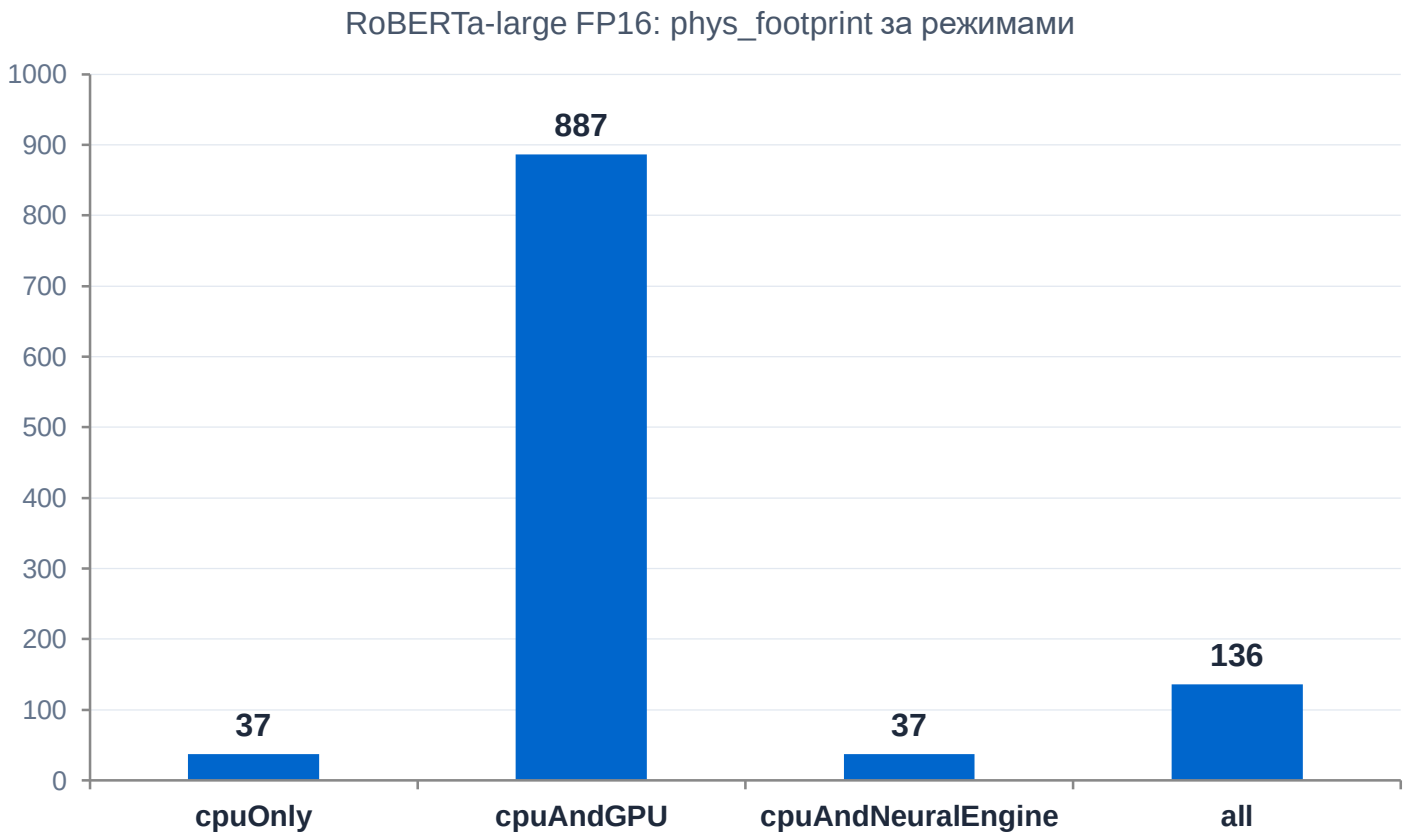
77 МБ

Чому саме клавіатура:

- Інтегрована в будь-яке поле введення
- Природний UX без зайвих дій
- Найжорсткіший ліміт пам'яті
- Якщо вписалися сюди — інші типи теж

Експеримент 2: класифікатор RoBERTa — ANE працює

Попередній експеримент для підтвердження гіпотези про виконання поза процесом



>20×

різниця phys_footprint між GPU (887 МБ) і ANE (37 МБ)

Палетизація 4-біт

711 МБ → 178 МБ

точність 97% без змін

✓ Інтегровано в Keyboard Extension

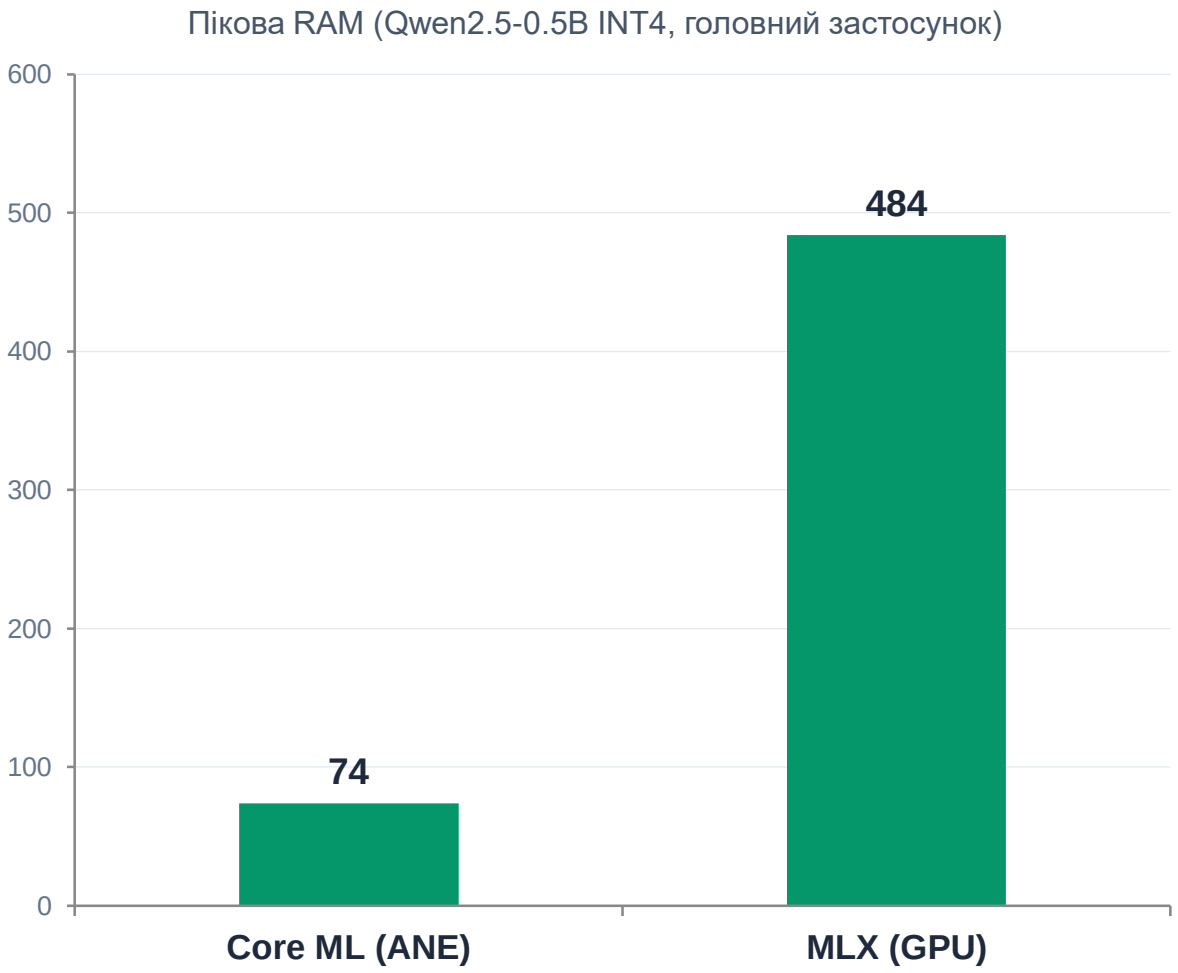
36 МБ phys_footprint у розширенні

в межах ліміту 77 МБ

Висновок: ANE дійсно виконує модель поза процесом. Гіпотеза підтверджена кількісно. Палетизація — оптимум для ANE.

Експеримент 3: Qwen2.5-0.5B — Core ML vs MLX

Виміри у головному застосунку для тієї ж моделі INT4. Хто з фреймворків теоретично сумісний з розширеннями?



ПОВНЕ ПОРІВНЯННЯ

Метрика	Core ML	MLX
Розмір, МБ	316	290
Пікова RAM, МБ	74	484
Завантаження, мс	109	1797
Перший токен, мс	546	118
Токенів/с	23	463

MLX: 484 МБ — не вписується навіть у Photo Editing (760 МБ для службових структур).

Core ML на ANE: 74 МБ — теоретично вписується навіть у Keyboard Extension (77 МБ).

→ *Core ML єдиний кандидат для подальшої спроби в розширенні.*

Експеримент 4: спроба в розширенні

Та сама конфігурація, що показала 74 МБ у головному застосунку, — у трьох типах розширень

Keyboard Extension

ліміт: 77 МБ

процес завершено

Action Extension

ліміт: 300 МБ

процес завершено

Photo Editing Extension

ліміт: 760 МБ

процес завершено

ДІАГНОСТИКА

phys_footprint швидко росте до розміру файлу моделі — ваги вантажаться в адресний простір розширення.

Найбільш ймовірне пояснення: XPC-доступ до системного демону ANE з розширень обмежений суворіше, ніж з основного застосунку. Core ML фолбекає на CPU, що матеріалізує ваги в процесі.

Цікаве: RoBERTa-класифікатор у тій самій клавіатурі працює нормально. Обмеження залежить від класу моделі.

Висновки та внесок роботи

НАУКОВИЙ ВНЕСОК

1. Виміряно фактичні ліміти phys_footprint для 4 типів iOS-розширень на A15 / iOS 26.4.
2. Кількісно підтверджено виконання моделей на ANE поза процесом для класифікаторів (>20× різниця).
3. Виявлено та задокументовано обмеження: ANE-оптимізація пам'яті недоступна для генеративних LLM у розширеннях, попри доступність для класифікаторів.
4. Показано несумісність MLX з лімітами розширень для будь-яких варіантів моделі.

ПРАКТИЧНІ РЕКОМЕНДАЦІЇ

Для класифікаційних задач

Core ML + ANE + 4-бітна палетизація працює у Custom Keyboard. Робоча конфігурація.

Для генеративних задач

Винести інференс LLM у головний застосунок, обмінюватися з розширенням через XPC або App Group.

Альтернатива

Енкодерні архітектури для шаблонних/класифікаційно-вибіркових задач замість повної генерації.

Перспективи

Очікування змін у програмному стеку iOS для повноцінного доступу розширень до ANE.

Поточна архітектура iOS не дозволяє повноцінного розгортання генеративних LLM у процесах-розширеннях — і це задокументований у роботі емпіричний факт.

Дякую за увагу.

Готовий до запитань.

Гібський Владислав Ігорович

Кваліфікаційна робота · НаУКМА · 2026