ДОДАТОК А

Графічний матеріал кваліфікаційної роботи

Харківський національний університет радіоелектроніки

Кваліфікаційна робота

«Алгоритми підвищення ефективності згорткових нейронних мереж»

Виконав :
ст. гр. КСМм-21-1
Пономаренко Р.Д.

Керівник :
проф. Міхаль О.П.

Мета та завдання кваліфікаційної роботи 2

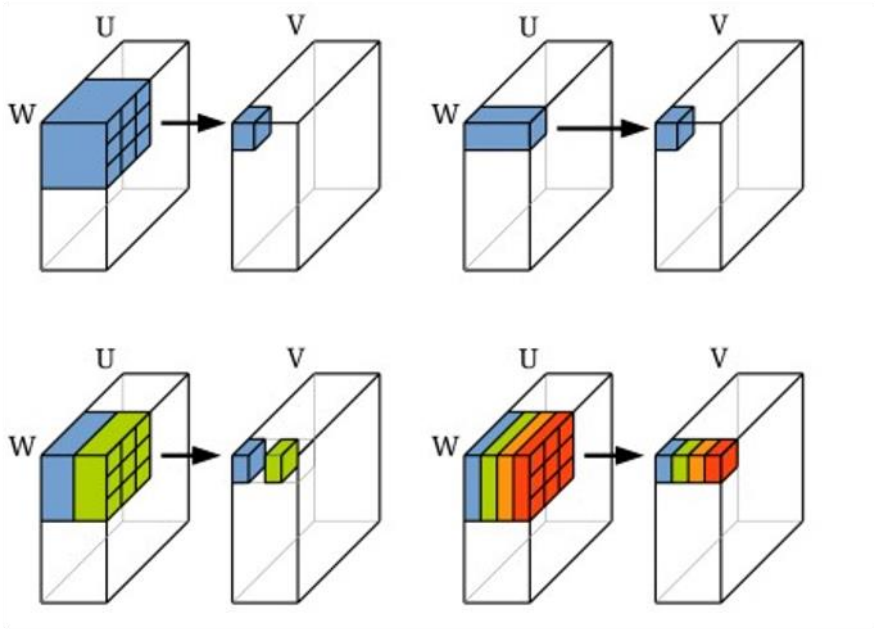
Мета кваліфікаційної роботи: аналіз методів та алгоритмів підвищення ефективності згорткових нейронних мереж.

Об'єкт дослідження: згорткові нейронні мережі.

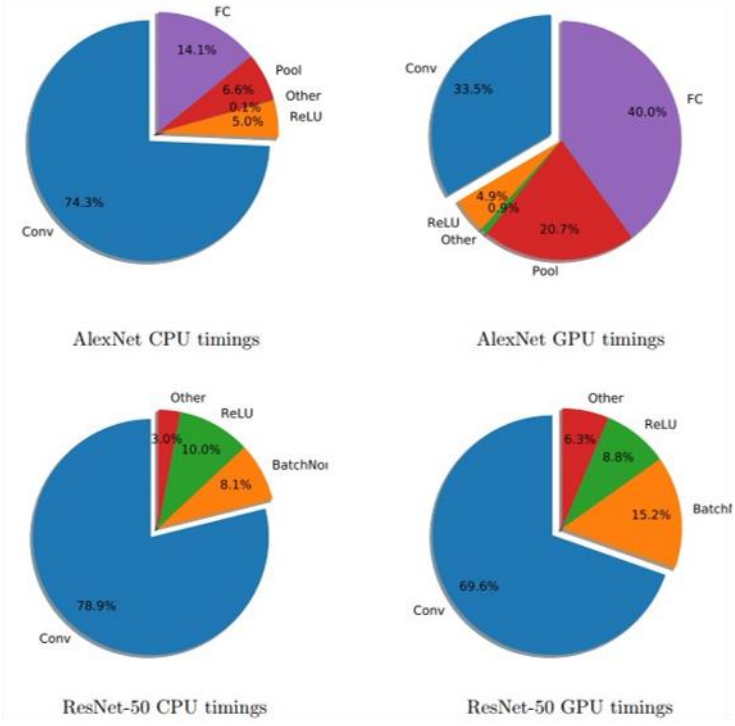
Завдання:

- ❑ аналіз популярних архітектур згорткових нейронних мереж та їх блоків;
- ❑ огляд існуючих алгоритмів підвищення ефективності згорткових нейронних мереж, а також методів тензорної декомпозиції;
- ❑ вибір наборів даних для проведення експериментів;
- ❑ порівняльний аналіз та реалізація розглянутих алгоритмів.

Варіанти згортки у сучасних розробках ЗНМ 3

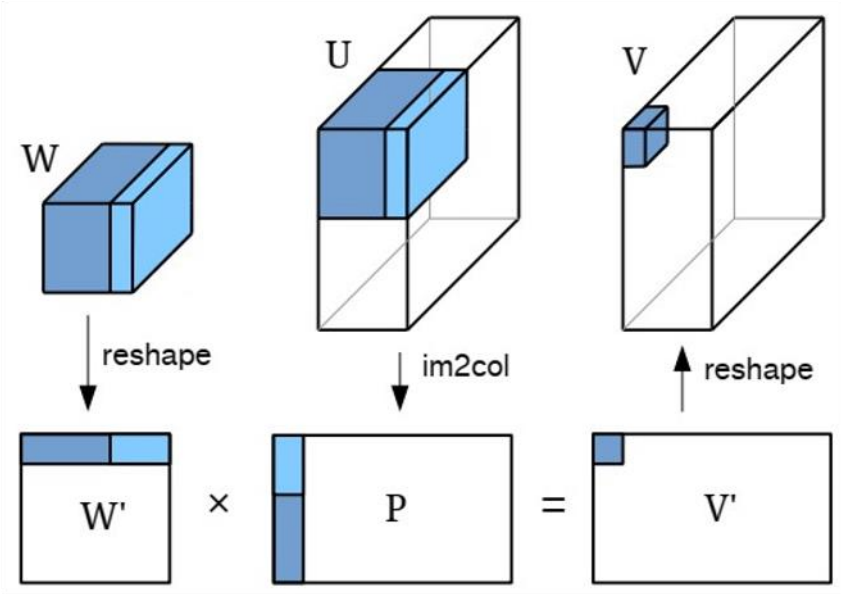


Таймінг різних рівнів для архітектур AlexNet та ResNet-50 4



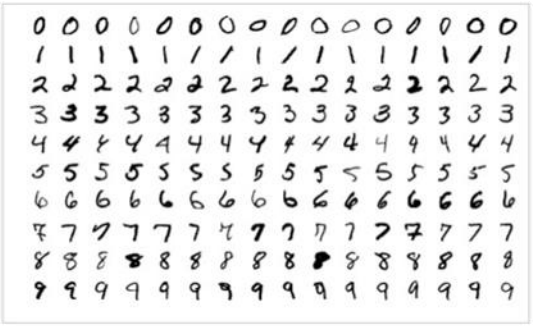
Зведення згортки

5

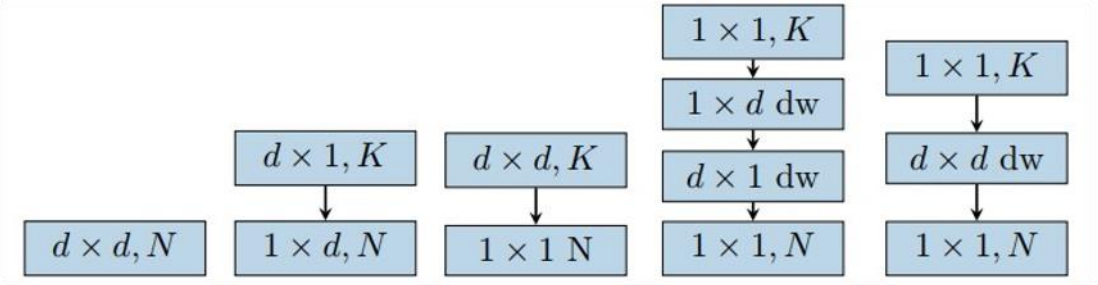


Набори даних

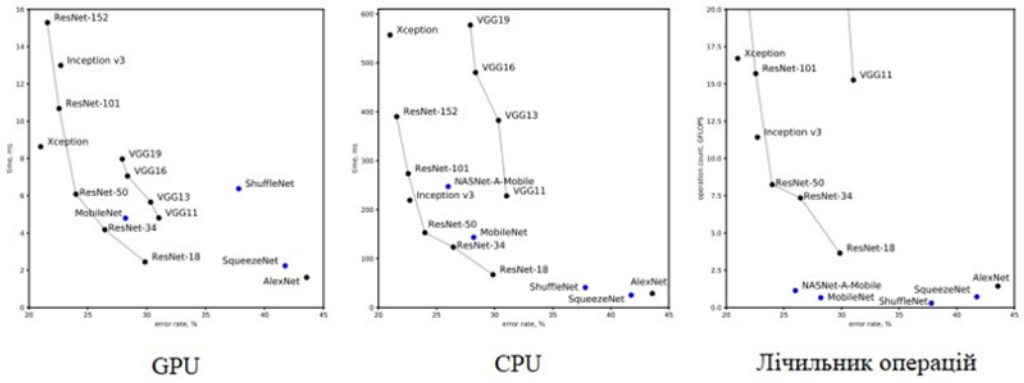
6



Блоки ЗНМ, які використовуються методами тензорної декомпозиції 7

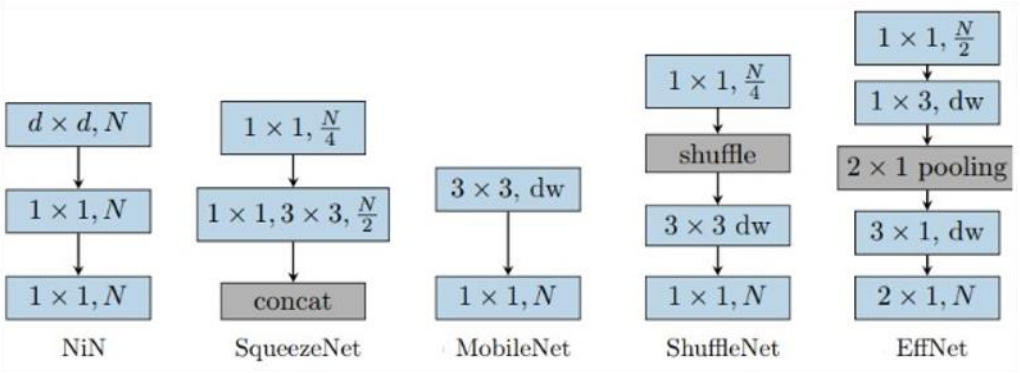


Діаграми точності ILSVRC та часу виводу для доступних архітектур ЗНМ 8



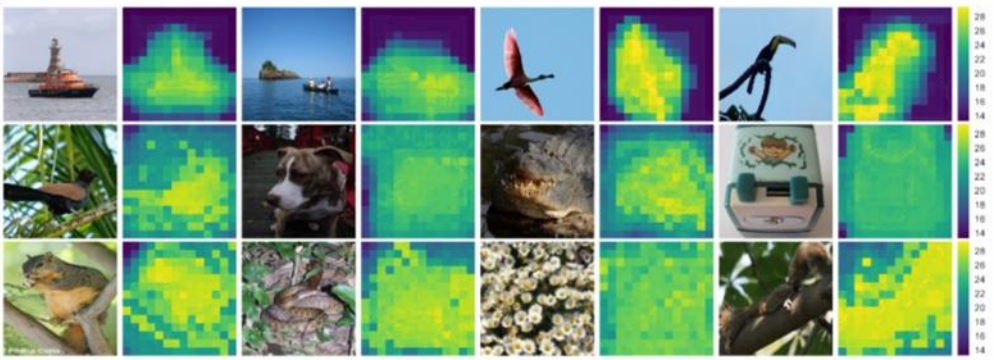
Послідовності згорткових шарів

9



10

Карти для перевірки ILSVRC



Алгоритм CP-декомпозиції

11

$$A(i,j) = \sum_{r=1}^{R} A_1(i,r) A_2(j,r), \quad i = \overline{1,n}, \quad j = \overline{1,m},$$

$$A(i_1, \ldots, i_d) = \sum_{r=1}^{R} A_1(i_1, r) \ldots A_d(i_d, r)$$

$$V(x,y,k) = \sum_{i=1}^{d} \sum_{j=1}^{d} \sum_{c=1}^{C} W(i,j,c,k)\, U(x+i, y+j, c)\,,$$

$$W(i,j,c,k) = \sum_{r=1}^{R} W^x(i,r)\, W^y(j,r)\, W^c(c,r)\, W^k(k,r)\,,$$

$$U^c(x,y,r) = \sum_{c=1}^{C} W^c(c,r)\, U(x,y,c)$$

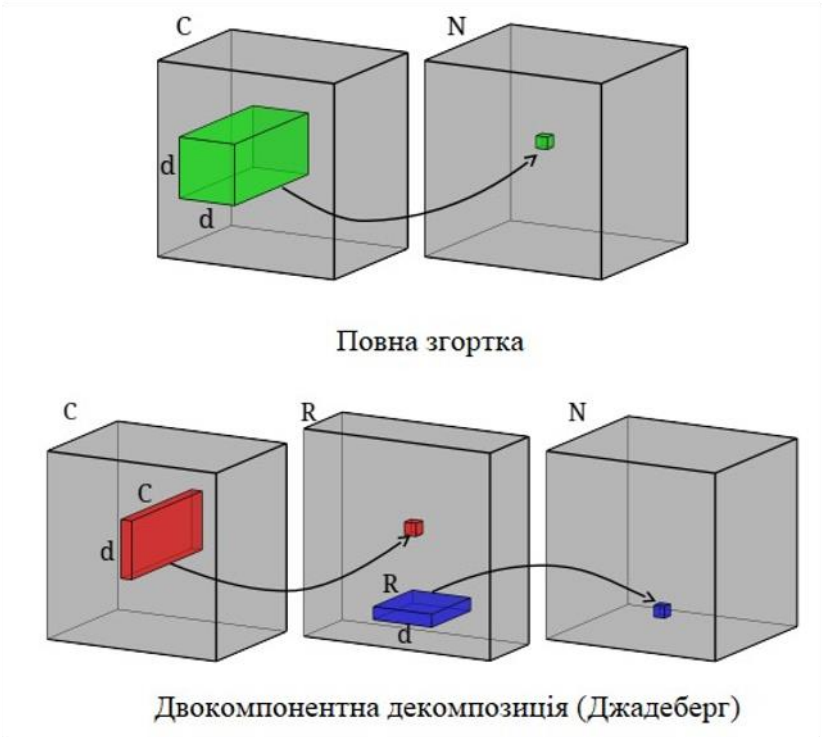
$$U^{cy}(x,y,r) = \sum_{j=1}^{d} W^y(j,r)\, U^s(i, y+j, r)$$

$$U^{cyx}(x,y,r) = \sum_{i=1}^{d} W^x(i,r)\, U^{cy}(x+i, y, r) \qquad V(x,y,k) = \sum_{r=1}^{R} W^k(k,r)\, U^{cyx}(x,y,r)\,,$$

$$V(x,y,k) = \sum_{r=1}^{R} W^k(k,r) \left(\sum_{i=1}^{d} W^x(i,r) \left(\sum_{j=1}^{d} W^y(j,r) \left(\sum_{c=1}^{C} W^c(c,r) U(x+i, y+j, c) \right) \right) \right)$$

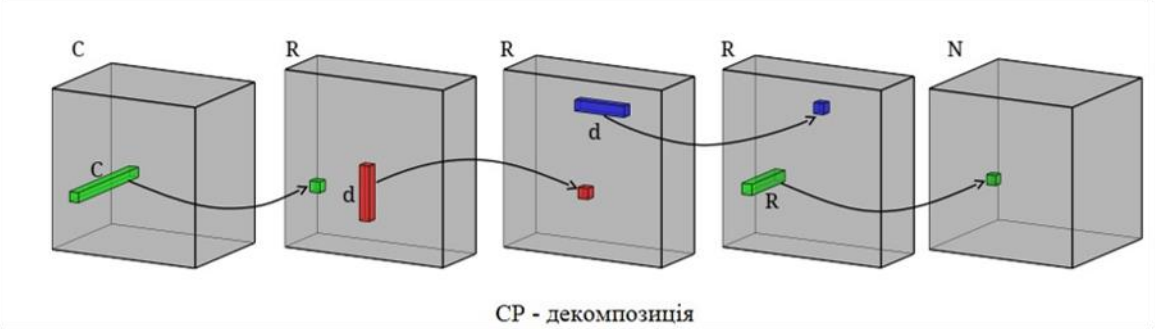
Тензорні розкладання для прискорення узагальненої згортки

12



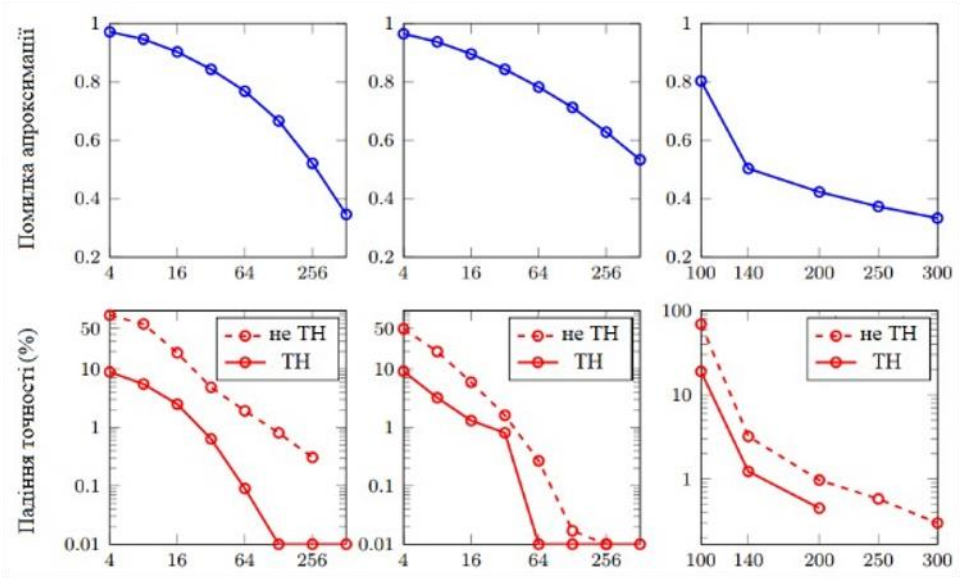
Тензорні розкладання для прискорення узагальненої згортки

13



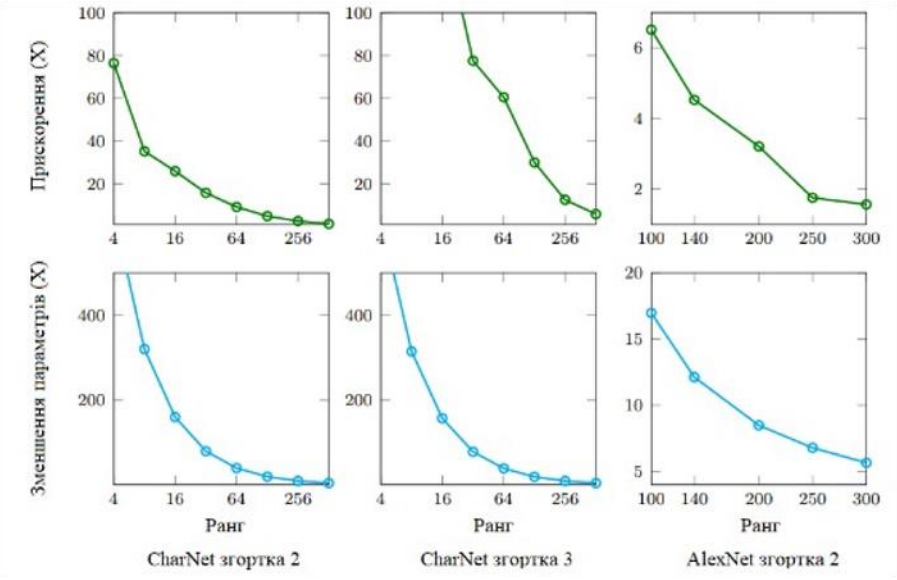
Результати експериментів

14



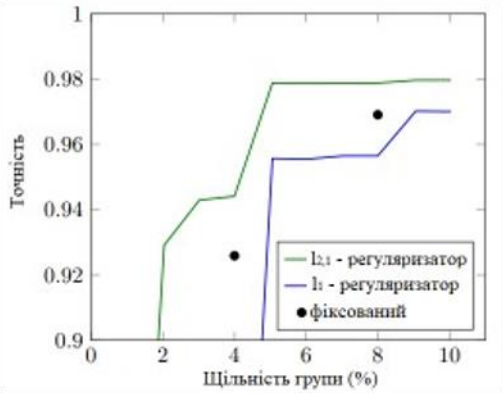
Результати експериментів

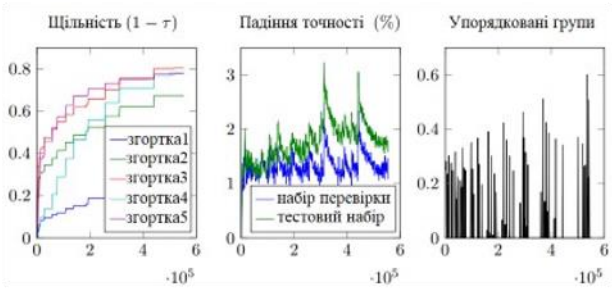
15



Результати експериментів

16





Висновки

17

Проведено аналіз методів та алгоритмів прискорення згорткових нейронних мереж. Вони поділяються на кілька груп: тензорні розклади, квантування, скорочення, навчання з вчителем, пошук ефективних архітектур та адаптивних моделей. Детально розглянуто алгоритми прискорення згортковиху нейронних мережах з низькоранговим CP-розкладом згорткових ваг. Реалізація методів базується на існуючих блоках ЗНМ, що дозволяє легко розгортати, і найголовніше, точніше налаштовувати моделі, хоча нестабільність CP-декомпозиції ускладнює процес тонкого налаштування. Експериментальні результати демонструють значне прискорення з мінімальним падінням точності для кількох архітектур.