ДОДАТОК А

Графічний матеріал кваліфікаційної роботи

Харківський національний університет радіоелектроніки

«Методи апаратного прискорення штучних нейронних мереж на FPGA»

Кваліфікаційна робота

Виконав:
ст. гр. КСМзм-22-1
Беляков А.А.

Керівник:
доц. Козлов Ю.В.

Мета та завдання кваліфікаційної роботи

2

Мета кваліфікаційної роботи: аналіз існуючих методів апаратного прискорення штучних нейронних мереж на FPGA.

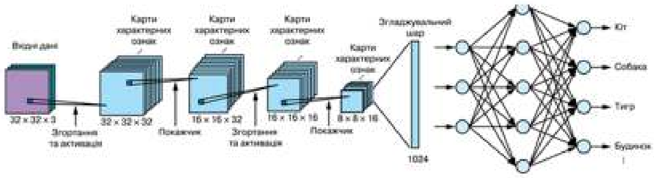
Об'єкт досліджень: штучні нейронні мережі, зокрема згорткові нейронні мережі.

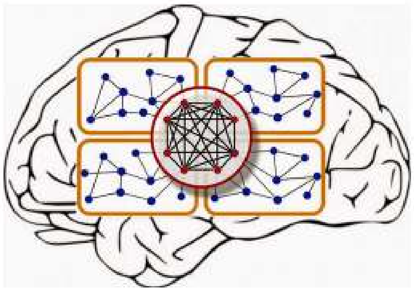
Завдання:

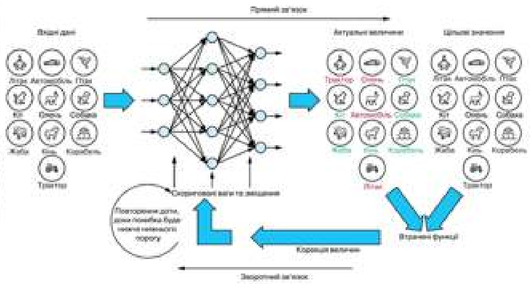
- аналіз топологій згорткових нейронних мереж;
- аналіз існуючих алгоритмів навчання згорткових нейронних мереж;
- розробка методики апаратного прискорення штучних нейронних мереж на FPGA;
- дослідження існуючих рішень апаратного прискорення;
- розробка інструменту генерації VHDL.

Штучні нейронні мережі. Навчання згорткових нейронних мереж з використанням CIFAR-10

3

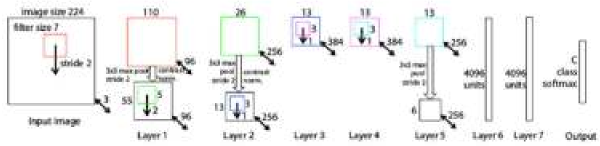


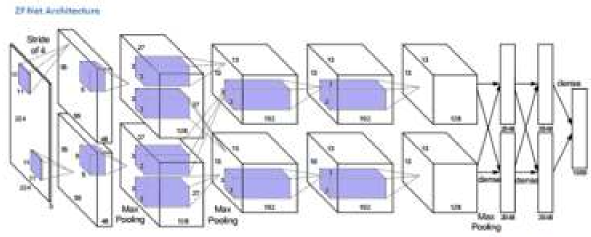


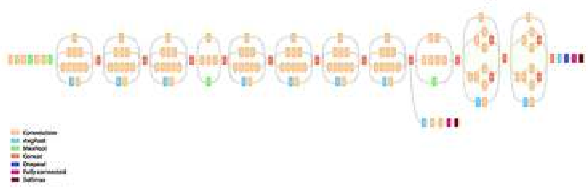


Топології згорткових нейронних мереж

4

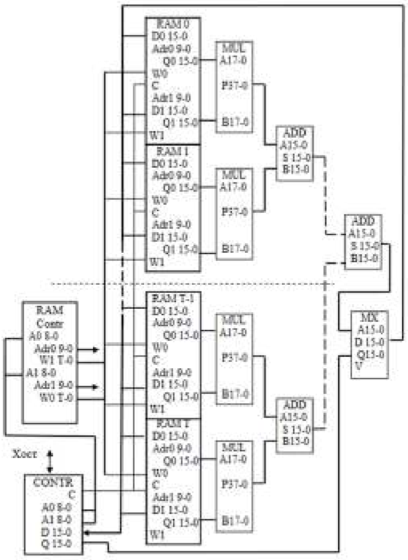






Огляд існуючих рішень. Структурна схема нейромережі на FPGA з паралельним обчисленням ваг

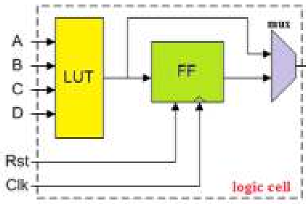
5

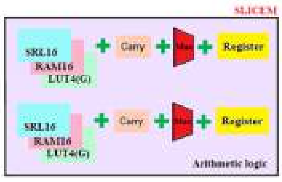


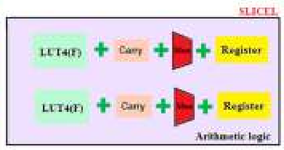
Розробка і побудова нейронних мереж у ПЛІС FPGA дозволяє значно прискорити процеси їх функціонування і навчання за рахунок використання можливості паралельної обробки від сотень до декількох тисяч обчислювальних потоків. Даний варіант структури нейронної мережі відрізняється високим коефіцієнтом використання кристала і можливістю застосування для реалізації згорткових нейронних мереж.

Спрощена архітектура FPGA. FPGA серії Xilinx Spartan 3E

6

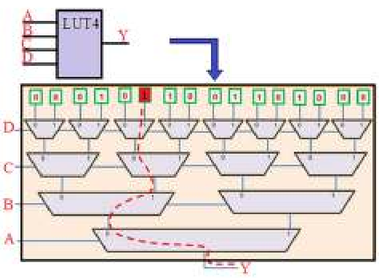


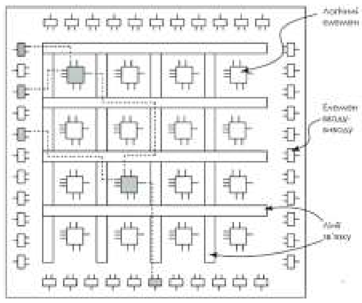




Truth Table

Inputs				Output
A	B	C	D	Y
0	0	0	0	0
0	0	0	1	0
0	0	1	0	0
0	0	1	1	1
0	1	0	0	0
0	1	0	1	1
0	1	1	0	1
0	1	1	1	0
1	0	0	0	0
1	0	0	1	1
1	0	1	0	1
1	0	1	1	0
1	1	0	0	1
1	1	0	1	0
1	1	1	0	0
1	1	1	1	0



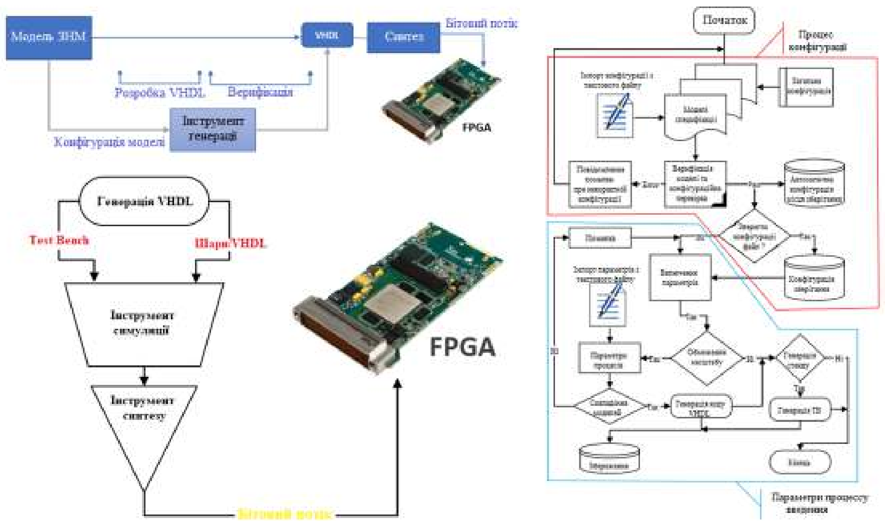


Методика апаратного прискорення 7

Методика апаратного прискорення ШНМ на FPGA з використанням інструменту генерації коду

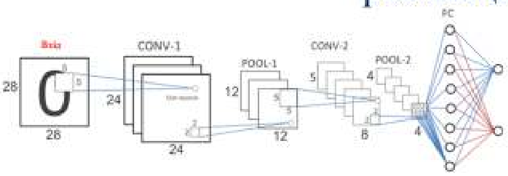
Зміст	Впровадження інструменту генерації VHDL на основі FPGA для реалізації згорткових нейронних мереж. Для розробки використовується мова програмування Java.
Призначення	Метод призначений для полегшення процесу апаратного прискорення моделей згорткових нейронних мереж з використанням FPGA шляхом параметризації реалізації цих моделей.
Вимоги	Інструмент має містити графічний інтерфейс користувача, за допомогою якого можливе налаштування своєї цільової моделі згорткової нейронної мережі, надавши специфікації моделі. Інструмент повинен бути оптимізован для створення модульної, масштабованої, реконфігурованої та паралельної реалізації моделей згорткових нейронних мереж.
Переваги	Використання методу значно зменшить час розробки, необхідний для впровадження згорткової нейронної мережі, також знащить недоліки, що виникають через складність розробки мов опису апаратних засобів, і має пом'якшити недостатню оптимізацію, викликану інструментами синтезу високого рівня.

Інструмент генерації VHDL. Процес впровадження моделі згорткової нейронної мережі 8

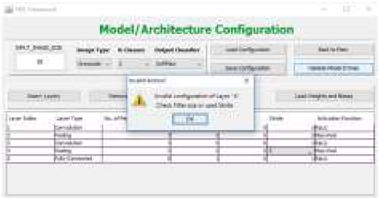


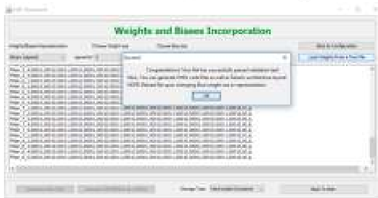
Тестова модель згорткової нейронної мережі та її реалізація

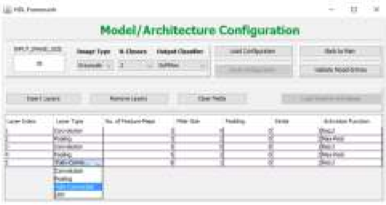
9

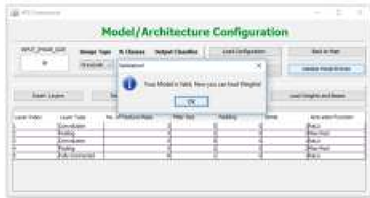


Цільова модель використовується для класифікації зображень рукописних 0 і 1. Вхідне зображення має відтінки сірого розміру 28 × 28. У першому згортковому шарі (Conv-1) витягуються три різні карти ознак шляхом згортання вхідного зображення з ядрами/фільтрами розміром 5 × 5. Крок згорткового фільтра дорівнює одиниці, до вхідного зображення не застосовується заповнення, а застосована функція активації.



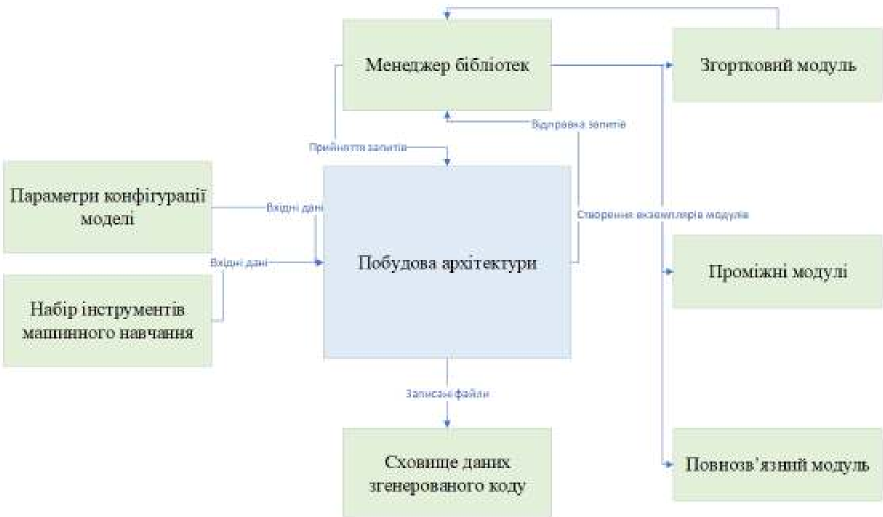






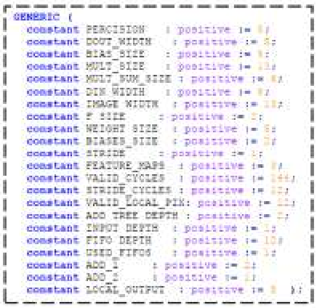
Процес генерації коду VHDL

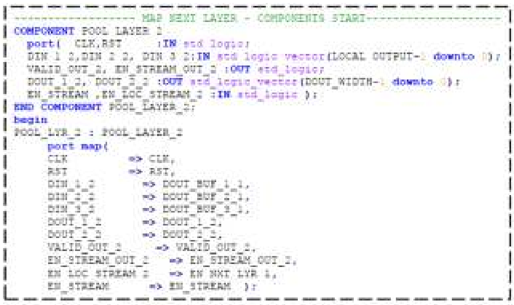
10



Приклади знімків згенерованого VHDL коду для першого рівня згортки

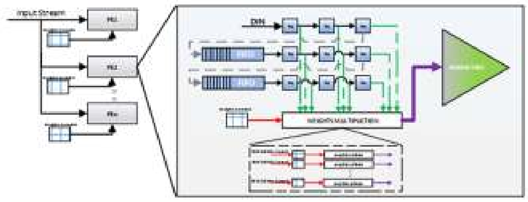
11

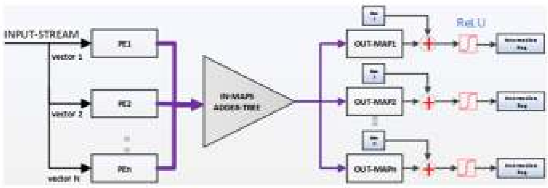


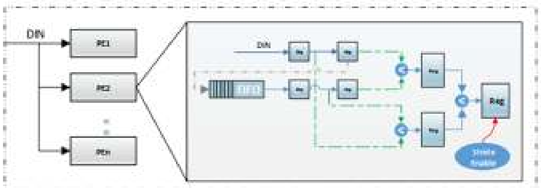


Результати роботи

12







Висновки

13

Проаналізовано методи апаратного прискорення штучних нейронних мереж на FPGA з використанням мови VHDL. Проведено аналіз топологій згорткових нейронних мереж; аналіз існуючих алгоритмів навчання згорткових нейронних мереж. Досліджено існуючі рішення апаратного прискорення.

Запропоновано методику апаратного прискорення ШНМ за допомогою FPGA. Розроблений інструмент генерації VHDL. Використання розробленої методики значно зменшить час розробки, необхідний для впровадження згорткової нейронної мережі, також мінімізує недоліки, що виникають через складність розробки мов опису апаратних засобів.

ДОДАТОК Б

Тези доповіді

МЕТОДИ АПАРАТНОГО ПРИСКОРЕННЯ ШТУЧНИХ НЕЙРОННИХ МЕРЕЖ

Беляков А.А., Козлов Ю.В.
Харківський національний університет радіоелектроніки, Харків, Україна

Нейронні мережі є одним з ключових елементів сучасних систем штучного інтелекту. З появою глибоких нейронних мереж (ГНМ)[1,2] та їх широким застосуванням у різних областях, таких як розпізнавання зображень, обробка мови та автономні транспортні засоби, стало актуальним питання про ефективність обчислень. Однак обчислювальна складність таких мереж часто потребує великих обчислювальних ресурсів. Поширеними підходами до прискорення нейронних мереж є використання графічних процесорів (ГП) та спеціалізованих інтегрованих схем (СІС). Завдяки великому числу паралельних ядер, ГП добре підходять для обчислень, що пов'язані з нейронними мережами. СІС же, у свою чергу, оптимізовані для конкретних завдань, таких як обробка глибоких мереж.

Метою доповіді є аналіз методів апаратного прискорення штучних нейронних мереж за допомогою традійних підходів, а також реалізація за допомогою Field-Programmable Gate Array (FPGA). FPGA набувають популярності як інструмент для апаратного прискорення нейронних мереж завдяки їх гнучкості та ефективності. Перевагами FPGA є: гнучкість, тобто FPGA можна перепрограмувати; енергоефективність, коли FPGA в порівнянні з GPU, може виконувати обчислення при меншому споживанні енергії; паралелізм, коли FPGA може виконувати численні операції одночасно завдяки архітектурі, що підтримує паралелізм. Хоча FPGA пропонують багато переваг, існують певні виклики такі як складність, вартість та обмежена кількість ресурсів. Розробка для FPGA може бути більш складною порівняно з традиційним програмуванням. FPGA можуть бути дорожчими за інші рішення, особливо на початкових етапах розробки. Також є певні обмеження на кількість логічних блоків та пам'ять на FPGA. Можливо, вирішенням даної проблеми може бути апаратне присорення на Tensor Processing Units (TPU) - це інший вид спеціалізованого апаратного прискорення, розроблений Google спеціально для глибокого навчання. Це є предметом подальших досліджень. Апаратне прискорення є важливим елементом в розробці та розгортанні ефективних нейронних мереж. З різноманітністю доступних технологій, таких як ГП, СІС, FPGA та TPU, дослідники та інженери можуть вибирати найкращий підхід в залежності від конкретних потреб та обмежень.

Список літератури

1. A. Coates, A. Arbor, and A. Y. Ng, “An Analysis of Single-Layer Networks in Unsupervised Feature Learning,” Aistats 2011, pp. 215–223, 2011.
2. Q. V Le, A. Coates, B. Prochnow, and A. Y. Ng, “On Optimization Methods for Deep Learning,” Proc. 28th Int. Conf. Mach. Learn., pp. 265–272, 2011.

Рисунок Б.1 – Наукова публікація