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.