УДК 004.54 (045)
Бортник К.Я.
Луцький національний технічний університет

СТРУКТУРА СИСТЕМИ БАГАТОШАРОВОГО ПЕРЦЕПТРОНУ ДЛЯ МОБІЛЬНИХ ПРИСТРОЇВ

Бортник К.Я. Структура системи багатошарового перцептрону для мобільних пристроїв. Розглянуто структуру системи на чіпі (SoC) багатошарового перцептрону для мобільних пристроїв. Система придатна для пристроїв, які мають потребу у реалізації функції розпізнавання образів і які мають обмежені обчислювальні можливості. Проведено проектування системи, а також її верифікація на спеціалізованій макетній платформі. Для розроблення програмного забезпечення використано мову моделювання та синтезу програмного забезпечення VHDL. Доведено, що перенесення нейронних обчислень з центральних процесорів загального призначення на апаратні платформи на базі програмованих логічних інтегральних схем (FPGA) значно прискорює роботу алгоритмів розпізнавання.
Ключові слова: програмована логічна інтегральна схема (FPGA), система на чіпі (SoC), багатошаровий перцептрон (MLP), штучна нейронна мережа (ANN).

Бортник К.Я. Структура системы многослойного персептрона для мобильных устройств. Рассмотрена структура системы на чипе (SoC) многослойного персептрона для мобильных устройств. Система пригодна для устройств, которые нуждаются в реализации функции распознавания образов и имеющих ограниченные вычислительные возможности. Проведено проектирование системы, а также ее верификация на специализированной макетной платформе. Для разработки программного обеспечения использован язык моделирования и синтеза программного обеспечения VHDL. Доказано, что перенос нейронных вычислений от центральных процессоров общего назначения на аппаратные платформы на базе программируемых логических интегральных схем (FPGA) значительно ускоряет работу алгоритмов распознавания.
Ключевые слова: программируемая логическая интегральная схема (FPGA), система на чипе (SoC), многослойный персептрон (MLP), искусственная нейронная сеть (ANN).

Bortnyk Ka. Structure of the system based on Multi-Layer Perceptron for Smart Devices. The structure of the system on chip (SoC) multilayer perceptron for mobile devices was observed. The system is suitable for devices that need to implement the function of pattern recognition and which have limited computing power. A system design and its verification made on a specialized prototyping platform. For software development language used simulation and synthesis software suite VHDL. It is considered that the transfer of neural computation from the CPU of general purposes to hardware platform based on field-programmable gate array (FPGA) significantly speeds up recognition algorithms.
Keywords: field-programmable gate array (FPGA), system on chip (SoC), Multilayer Perceptron (MLP), artificial neural network (ANN).

Актуальність проблеми. Попит на «розумні» пристрої споживчої електроніки зростає. Це мотивується широким використанням недорогих вбудованих електронних пристроїв різноманітного призначення. Окрім того, бажано, щоб електронні пристрої мали здатність відчувати і розуміти їх оточення та адаптувати їхні сервіси відповідно до контексту.

Штучні нейронні мережі (ANN) можуть бути обрані для цієї мети у першу чергу, у зв'язку з їх широким спектром застосовності.

Багатошаровий перцептрон (MLP) є однією з найбільш часто використовуваних ANN через його здатність моделювати нелінійні системи і встановлювати межі нелінійних рішень в задачах класифікації, таких як оптичне розпізнавання символів (OCR), інтелектуальний аналіз даних і обробки/ розпізнавання зображень.

Однак, оскільки MLP вимагає надзвичайно високої пропускної здатності, його обчислювальна складність вкрай небажана для виконання операцій в режимі реального часу, особливо для вбудованих пристроїв, які мають обмеження у своїх можливостях обробки даних. Привабливе вирішення цього полягає в розробці спеціалізованих апаратних засобів для прискорення MLP.

Апаратна реалізація MLP була нагальною темою протягом багатьох років, в основному у питаннях точності, необхідного простору і швидкості обробки. Різні реалізації обладнання для MLP були успішними, наприклад, метод проектування спеціалізованих інтегральних схем. Тим не менш, повна апаратна реалізація не є ефективною з точки зору вартості пристрою та складності його реалізації.

Останнім часом активно досліджується перенастроювана обчислювальна парадигма, в рамках якої розроблено ряд програмованих логічних інтегральних схем (FPGA). Але хоча кілька

апаратних реалізацій з використанням FPGA вже були запропоновані, апаратна реалізація в MLP, як і раніше, залишається складним завданням для вбудованих додатків.

Оскільки, різні методи преобробки і постобробки можуть бути суміщені з MLP в реальних додатках, система повинна настроюватись відповідно до додатків. Крім того, існує важлива вимога до проектування апаратного забезпечення таким чином, щоб внесення змін в структуру мережі не приводило до необхідності апаратної модернізації.

Ці проблеми можуть бути подолані за допомогою комбінованого програмного/апаратного методу проектування. Цей метод здійснюється шляхом аналізу різних частин алгоритму та реалізації його найбільш вимогливих ланок на швидких апаратних пристроях. Служити цьому може так звана система на чіпі (SoC), що складається з мікропроцесора та апаратних прискорювачів на базі FPGA, що може значно прискорити роботу, зберігаючи значною мірою гнучкість суто програмних моделей.

У цій статті описано архітектуру MLP-SoC смарт-додатків для вбудованих пристроїв. У SoC можна вносити зміни в структуру мережі і додатків без модифікації апаратних засобів.

Завдання з реалізації MLP-SoC. Нашою метою є розробка MLP-SoC, яка може бути використана для вбудованих додатків. У ході реалізації MLP-SoC на макетній платформі представлення даних, точність і апаратні компоненти відіграють важливу роль у виборі проектних рішень.

Макетна платформа. Для розробки та перевірки SoC найбільш доцільно використати основану на FPGA макетну платформу XILINX X2CV8000 , яка стала популярною саме через можливість швидкого макетування і перевірки вбудованих додатків.

Макетування всього проекту цільової SoC в FPGA дає точне і швидке уявлення про майбутній пристрій.

Звичайно, деякі основні компоненти, включаючи центральний процесор, системи шин і внутрішніх комутованих блоків, вибираються для проектування реальної платформи на виробництві. Одначе, для розробки і тестування алгоритмів додатку достатньо вбудованих у макетну платформу. Наприклад, мікропроцесор LEON 2 вбудований в FPGA, шина AHB/APB AMBA, призначена для зв'язку між внутрішніми компонентами, також вже вбудована в FPGA.

На додаток до мікросхеми FPGA, макетна платформа пропонує пам'ять на основі SDRAM чіпів розміром 128 МБ і флеш-пам'ять розміром 8 Мб. Блок пам'яті SDRAM використовується як оперативний запам'ятовучий пристрій, у той час, як на базі флеш-пам'яті організовано ПЗП для зберігання програмного забезпечення. На рисунку 1 зображено макетну платформу, що має досить компактний розмір 112 на 129мм.

Рис. 1. Макетна платформа XILINX X2CV8000

MLP для обробки/розпізнавання зображень. MLP для обробки/розпізнавання зображення складається з вузлів обробки даних, розташованих шарами.

Як правило, він вимагає трьох або більше шарів вузлів обробки: вхідний шар, один або декілька внутрішніх шарів та вихідний прошарок. Кожен вузол обробки в одному конкретному шарі повністю або частково пов'язано з кожним вузлом шару, що розташований вище і нижче за

нього. Зважені з'єднання визначають поведінку в мережі і налаштовуються під час тренування мережі за допомогою спеціального алгоритму зворотного поширення.

Під час розпізнавання, вхідний вектор подається на вхідного шар. Для наступних шарів, вхід в кожен вузол є сумою скалярних добутків елементів вхідного вектора з відповідними їм ваговими коефіцієнтами:

$$sum_i = \sum_j w_{ij} \bullet out_j ,$$

де w_{ij} - вага підключення вузла до вузла *JI*, а out_j - вихід з вузла *J*.

Вихід вузла *I*, це $out_j = F(sum_i)$, який потім розсилається всім вузлам в наступному шарі. Ця операція повторюється для всіх шарів мережі доти, поки вихідний шар не буде досягнуто, де і обчислюється вихідний вектор. F – це функція активації кожного вузла. У якості функції активації найчастіше використовується сигмовидна функція або функція гіперболічного тангенса. У таблиці 1 показані апаратні обмеження для цільової SoC.

Таблиця 1.

Апаратні константи цільової SoC

Назва параметру	Значення
Максимальне число вхідних вузлів	1000
Максимальне число внутрішніх вузлів та шарів	128/2
Точність вагового коефіцієнта	12 біт (знаковий)
Точність виходу функції активації	9 біт (знаковий)
Діапазон вхідних даних	0 … 255
Діапазон вихідних даних	-255 … 255

Використання типу даних з плаваючою розрядністю (ваги, входи, виходи) в нейронній мережі може бути непрактичним для вбудованих апаратних засобів, тому варто використати тип даних з фіксованою розрядністю для значень ваг, входів і виходів.

Незнакові 8 бітні використовуються для представлення вхідних значень, у той час, як знакові 9 бітні використовуються для вихідних значень функції активації, бо вони можуть бути від'ємними. Коефіцієнти ваг зберігаються в таблиці ваг за допомогою знакових 12 бітних з фіксованою розрядністю.

Безпосередня апаратна реалізація певної функції активації не відповідає завданню, оскільки пристрій повинен мати можливість зміни конфігурації. Тому слід використати таблицю відповідності зберігання вихідних значень для визначення функцій активації. При використанні цього методу, кілька різних функцій активації можуть бути реалізовані апаратно.

Структура MLP-SoC. Рис. 2. показує загальну структуру МЛП-SoC. Вона включає в себе LEON 2 (основний процесор), MLP співпроцесор (нейронних обчислень), контролер пам'яті SDRAM, інтерфейс камери і шину AHB/APB AMBA. Всі ці компоненти об'єднані в FPGA макетної платформи.

LEON 2 являє собою 32-розрядний RISC-процесор сумісний з архітектурою SPARC V8. Він легко програмується і, таким чином, дуже підходить для SoC.

Окрім того, програмне забезпечення, написане мовою С, може бути безпосередньо виконане на ядрі LEON 2 з використанням засобу крос-трансляції.

Програмування ядра LEON 2 (рис. 3.) може бути здійснено з використанням відкритого транслятора мови VHDL в код FPGA макетної платформи. Контролер інтерфейсу камери і схема вводу/виводу I2C здатні обробляти декілька датчиків зображення з використанням їх фіксованої логіки.

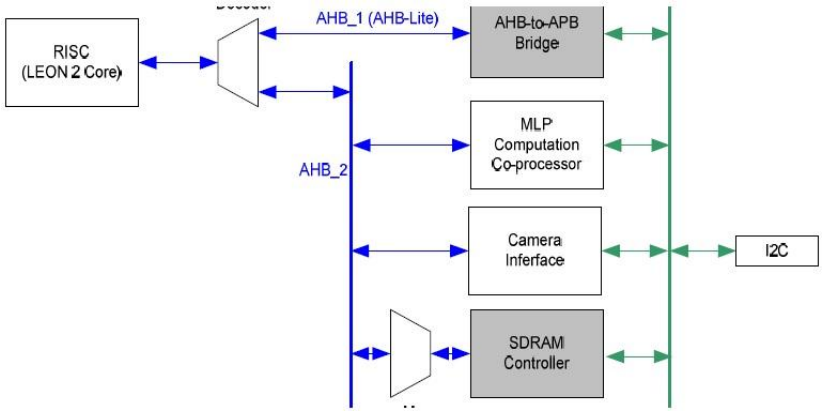


Рис. 2. Архітектурний огляд MLP-SoC

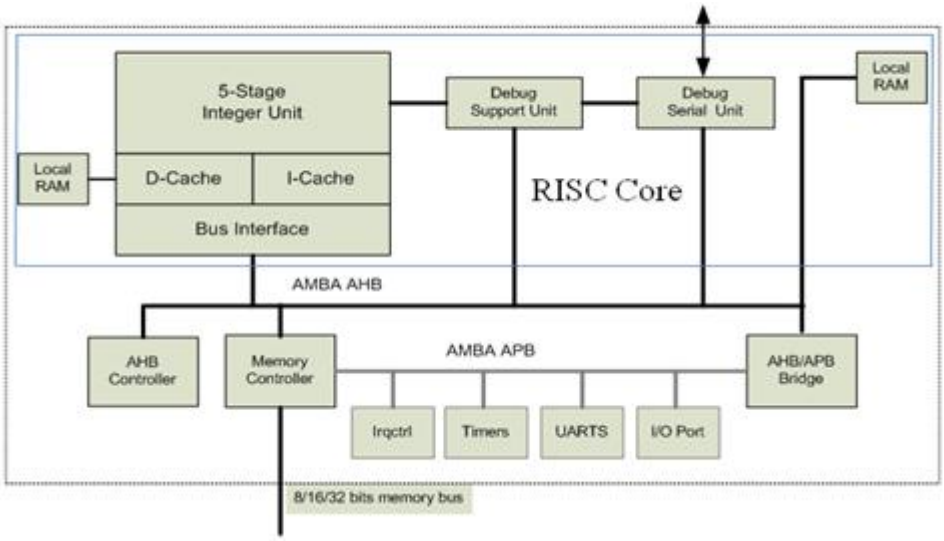


Рис. 3. Блок-схема процесора LEON 2

MLP співпроцесор На рис. 4 показана архітектура вбудованого MLP співпроцесора, основним призначенням якого є обчислення нейронів. Як видно на малюнку, співпроцесор складається з двох основних частин - блок хост-інтерфейсу для доступу до пам'яті та інтерфейс шини і MLP блок для нейронних обчислень.

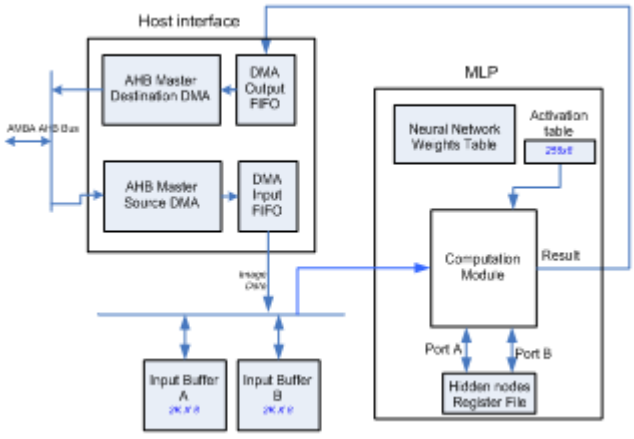


Рис. 4. Огляд архітектури MLP співпроцесора

Блок хост-інтерфейсу несе відповідальність за обслуговування шини між MLP співпроцесором та іншими контролерами.

Він складається з двох блоків прямого доступу до пам'яті (DMA), вхідного DMA і вихідного DMA. Вхідний DMA приймає блок даних із зовнішньої пам'яті і зберігає його у вхідних буферах (2K x 8 бітів). Два буфери завжди готові для прийому даних, а інші зберігають дані для подальших обчислень. Після прийому даних блок хост-інтерфейсу посилає сигнал до блоку MLP, щоб почати обчислення задачі для поточного вхідного сигналу. Коли обчислення завершується, вихідний DMA блок передає згенеровані блоки даних з МЛП блоку в зовнішню пам'ять. Використання DMA потоку даних є корисним, коли розмір і кількість вхідних/вихідних даних досить значні.

Для того, щоб врахувати обмеження, описані в таблиці 1, блок MLP складається зі сховищ даних і обчислювального модуля. Є три різних постійних блоків пам'яті: таблиця функцій активації , файл регістру внутрішніх вузлів і таблиця ваг.

Функція активації, наприклад, сигмовидна або гіпертангенціальна, можуть бути реалізовані в таблиці функцій активації без зміни апаратних засобів.

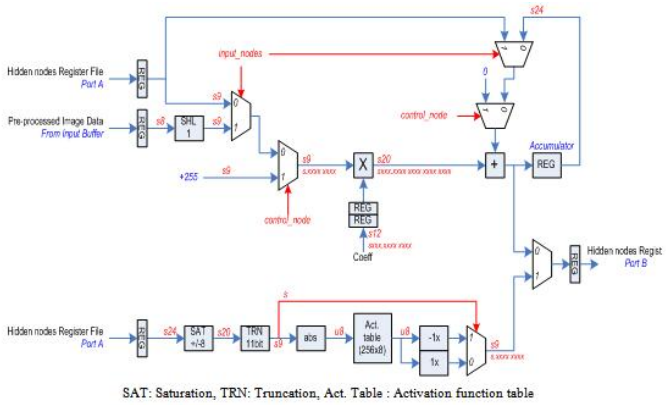
Таблиця ваг містить 128K * 19 біт, 12 біт використовуються для збереження значення ваги і 7 біт використовуються для збереження індексу внутрішнього вузла.

Файл регістру внутрішніх вузлів містить 128 * 24 біт для зберігання перехідних результатів внутрішніх вузлів.

Рис. 5. показує схему обчислювального модуля. Модуль обчислення отримує вхідні сигнали з вхідного буфера і обчислює значення активації всіх вузлів шарів послідовно до обчислення значенн вихідних вузлів. Потім він відправляє вихідні значення для блоку хост-інтерфейсу для збереження їх в пам'яті SRAM. Рис. 5 також демонструє точність вбудованої логіки.

Програмна реалізація МЛП-SoC може бути повністю синтезована на основі VHDL моделі і передані в FPGA (XILINX X2CV8000) макетної платформи.

Ця архітектура МЛП-SoC забезпечує швидку обробку нейронних зв'язків і трансферних функцій, і добре підходить для нейронних моделей MLP-типу. Тактова частота FPGA 30 МГц. Цей тактовий сигнал подається в усі компоненти макетної платформи.



SAT: Saturation, TRN: Truncation, Act. Table : Activation function table

Рис. 5. Схема обчислювального модуля MLP співпроцесора

Приклад додатку: Система розпізнавання символів (OCR). OCR процес, за допомогою якого комп'ютер переводить оцифровані зображення символів у текст. Ця система є основою для багатьох різних типів вбудованих додатків, таких як портативні перекладачі, електронні словники і щоденники. Алгоритм цільової системи розпізнавання складається з трьох основних етапів, як показано на рис. 6.

По-перше, отримання зображення зі сканера МКМ MT9V112, підключеного до інтерфейсу камери.

По-друге, попередньої обробки, яка виконується для того, щоб сегментувати зображення на окремі символи, використовуючи метод гістограм. Виділені символи перетворюються в бінарні матриці зображень (0 або 255). Тоді проводиться нормування похилих і підгонка розміру до отримання вхідного зображення MLP розміром 30x24 (в пікселях).

Оскільки структура нейронної мережі (число вузлів, функції активації) може змінюватися для конкретного додатка з метою підвищення продуктивності, а SoC для MLP повинна прийняти ці зміни без зміни конфігурації апаратної платформи, то, щоб показати здатність MLP_SoC до

переналаштування, слід побудувати два MLP на тій же архітектурі. Таблиця 2 показує конфігурації реалізованих MLP.

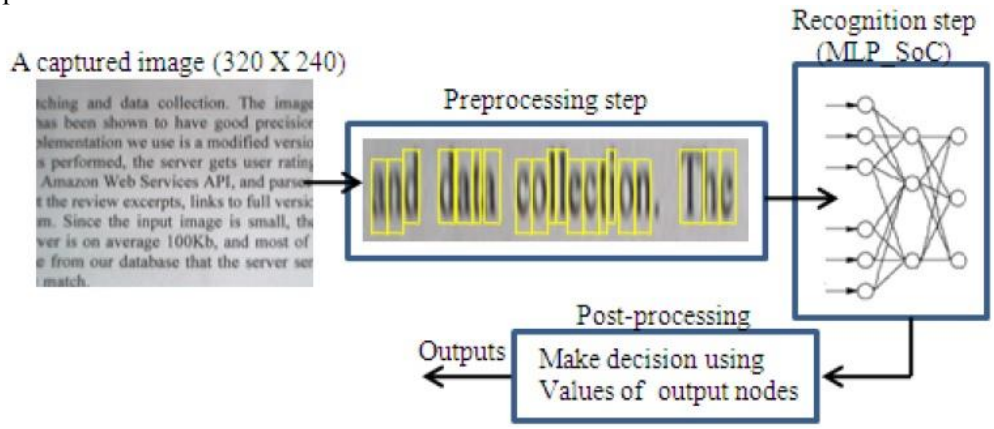


Рис. 6. Потік обробки реалізованої системи OCR

Оскільки структура нейронної мережі, число вузлів, функції активації, може змінюватися для конкретного додатка з метою підвищення продуктивності, A SoC для MLP повинні прийняти ці зміни без зміни апаратної платформи. То щоб показати здатність MLP_SoC до переналаштування, слід побудувати два MLP на тій же архітектурі. Таблиця 2 показує конфігурації реалізованих MLP. Кількість внутрішніх вузлів кожного MLP вибирається випадковим чином (3 ~ 5% від числа вхідних вузлів) та фіксується під час тренування і тестового розпізнавання.

Таблиця 2

Конфігурації MLP мереж

Назва параметру	**Версія 1**	**Версія 2**
Вузлів вхідних/внутрішніх/вихідних	720/24/26	720/32/26
Тип функції активації	Гіперболічно-тангенціальна	Сигмоїдна
Тип комутації вузлів	Повний	Частковий
Тип алгоритму навчання	Зворотнє поширення	Зворотнє поширення
Відсоток розпізнавання	94%	98%

Інше важливе питання для оцінки MLP-SoC є швидкість розпізнавання. Слід перевірити необхідний час роботи кожного модуля системи розпізнавання на MLP. У якості тестового зразка був взятий документ розміром 320X240 (пікселів), що містить 260 символів.

У таблиці 3 наведені час всіх модулів, затрачений для виконання цього завдання.

Таким чином, система розпізнавання здатна обробляти майже 43 символи в секунду. MLP обчислення займає всього 3.9 сек, в той час як програмна реалізація цільової MLP займає аж 869 секунд під час виконання на процесорі LEON 2.

Таблиця 3

Швидкість кожного модуля обробки для системи оптичного розпізнавання символів

Стадія алгоритму	**Модуль**	**Час**
Сегментування символів	Програма на LEON 2	690 мсек
Підготовка і нормалізація	Програма на LEON 2	1240 мсек
Нейронна мережа	Апаратні засоби	3980 мсек
Рішення і запис результату	Програма на LEON 2	120 мсек
Всього		6030 мсек

Цей результат в основному досягнутий за рахунок обчислення нейронної мережі MLP співпроцесором, що прискорило процес у 223 разів порівняно до програмної реалізації.

Комерційні системи програмного забезпечення OCR, реалізовані для серверів або настільних комп'ютерів, які зазвичай мають більш високі апаратні можливості, такі як потужні процесори, не займають багато часу для розпізнавання. Однак, мобільні пристрої часто мають обмеження у своїх апаратних можливостях через вимоги до обмеження споживання електроенергії.

Таким чином, апаратне прискорення є кращим рішенням для реалізації функцій розпізнавання образів у пристроях з обмеженими обчислювальними можливостями.

Висновки. У цій статті була запропонована архітектура MLP-SoC придатна для малогабаритних інтелектуальних пристроїв. Архітектура була випробувана і перевірена за допомогою макетної платформи FPGA. Платформа дає можливість без зміни існуючого обладнання побудувати різні варіанти прикладних систем шляхом реконфігурації SoC. Отже, MLP-SoC доцільно використати для мобільних пристроїв, які потребують функції розпізнавання образів.

1. Иванов А.И. Подсознание искусственного интеллекта:программирование автоматов нейросетевой биометрии языком их обучения. Электронная книга издательства ОАО "ПНИЭИ", 2012. —125 с.
2. E. M. Oritigosa, A. Canas, E. Ros, P. M. Ortigosa, S. Mota and J. Diaz, "Hardware description of multi-layer perceptrons with different abstraction levels," Microprocessors and Microsystems, vol. 30, pp. 435 – 444, 2006.
3. S. Vitabile, V. Conti, F. Gennaro and F. Sorbello, "Efficient MLP Digital Implementation on FPGA," Proceedings of the 8th Euromicro conference on DSD, 2005.
4. Новіков О., Кащенко С. Розпізнавання сервісів tcp/ip за допомогою нейронних мереж . Періодичний науково-технічний збірник " КПІ – 1, 2000 р., С 222-227.
5. Бонгард М. М. Проблема узнавания. — М. : Наука, 1967. — 320 с.
6. Керниган Б., Ритчи Д. Язык программирования Си. - СПб.: Невский Диалект,2001. – 352 с.
7. LEON2 processor user's manual, Gaisler Research, http://www.gaisler.com.