AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

Wydział Elektrotechniki, Automatyki Informatyki i Elektroniki

Katedra Elektroniki

Mariusz Kwiczala

Implementacja sieci neuronowych w układach FPGA

Autoreferat rozprawy doktorskiej

Promotor:

prof. dr hab. inż. Kazimierz Wiatr

Kraków 2009

1 Wstęp

Celem prezentowanej pracy jest sprzętowa implementacja sieci neuronowych w układach FPGA. Implementowana sieć ma służyć do przetwarzania obrazu cyfrowego. Ze względu na realizowane algorytmy przetwarzania obrazów sieć będzie składać się z dużej liczby połączonych ze sobą neuronów, aby mogła realizować wyznaczone cele. Przyjęto, że do sprzętowej implementacji wykorzystane zostaną znane sieci neuronowe oraz implementowane sieci będą już nauczone, co znacznie uprości strukturę układu. Dzięki zastosowaniu sprzętowej implementacji możliwe będzie znaczne zwiększenie szybkości przetwarzania sygnałów wejściowych w porównaniu z rozwiązaniami programowymi.

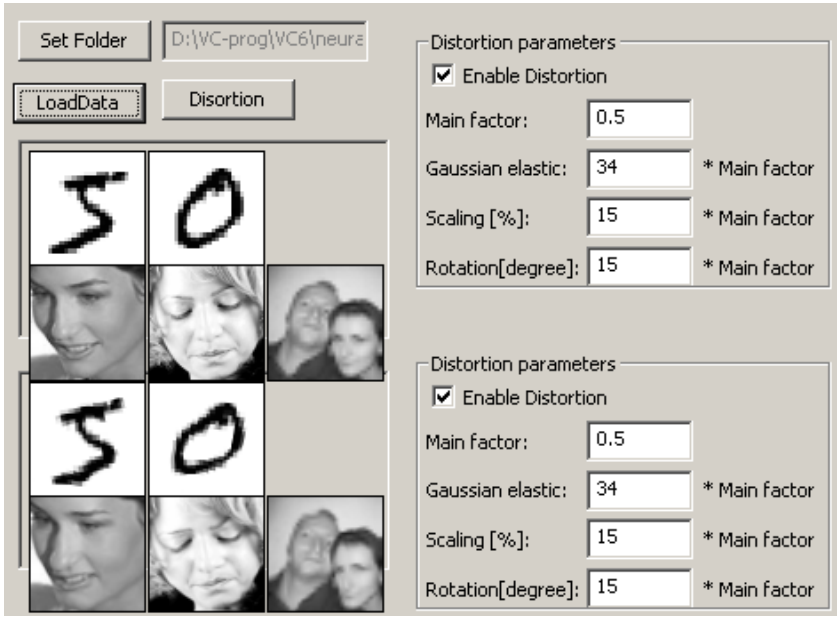
Ponieważ algorytmy przetwarzania obrazów i rozpoznawania obrazów są stosunkowo rozbudowane obliczeniowo, zaprojektowana sieć będzie złożona ze stosunkowo dużej liczby połączonych neuronów. Ograniczeniem jest wielkość zasobów sprzętowych, jakimi dysponujemy. Z drugiej strony czas przetwarzania obrazów powinien być możliwie krótki. Dzięki zrównolegleniu operacji przetwarzania w układach FPGA, możliwe jest wykorzystanie zalet równoległego przetwarzania, jakie niosą ze sobą sieci neuronowe.

2 Teza rozprawy

Sprzętowa implementacja sieci neuronowej w układach FPGA pozwala na znaczącą akcelerację obliczeń, w tym szczególnie przeznaczonych dla przetwarzania obrazów wizyjnych w czasie rzeczywistym.

3 Programowa implementacja sieci neuronowej

Pierwszy etap prac dotyczył programowej implementacji sieci neuronowej jednokierunkowej, wielowarstwowej testowanej na komputerze klasy PC. Programowa implementacja sieci neuronowej miała na celu wygenerowanie współczynników wagowych oraz weryfikację poprawności działania implementacji sieci neuronowej w układzie FPGA. Była także punktem odniesienia dla obliczania poziomu akceleracji obliczeń. Programowa sieć neuronowa jest uczona z wykorzystaniem zbiorów ręcznie pisanych cyfr oraz obrazów zawierających twarze lub pozbawionych ich wizerunków.



3.1 Zniekształcenia wzorca uczącego

Dane uczące zostały w trakcie badań celowo poddane zniekształceniu wzorca uczącego, który znacznie poprawia generalizację sieci neuronowej, a co się z tym wiąże poprawia skuteczność detekcji oraz rozpoznawania.

4 Sprzętowa implementacja sieci neuronowej

Implementacja sieci neuronowej w układzie FPGA pozwala na uzyskanie wielokrotnej akceleracji obliczeń w porównaniu z jej implementacją programową. Ze względu na ograniczenie zasobów sprzętowych układu FPGA oraz potrzeby zaimplementowania bardzo dużej sieci neuronowej bardzo ważne jest optymalne wykorzystanie dostępnych zasobów. Główne utrudnienia w implementacji sieci neuronowej to: reprezentacja danych, stopień zrównoleglenia operacji, cyfrowe mnożenie, duża liczba połączeń pomiędzy neuronami oraz funkcja aktywacji.

Zaimplementowano 3 rodzaje zrównoleglenia operacji: pełne zrównoleglenie NNFull, zrównoleglenie z wykorzystaniem arytmetyki rozproszonej NNDA, zrównoleglenie na poziomie węzłów NNNode. Każdy z rodzajów zrównoleglenia sieci neuronowej jest w pełni parametryzowany ze względu na liczbę neuronów, dokładność obliczeń i typ funkcji aktywacji.

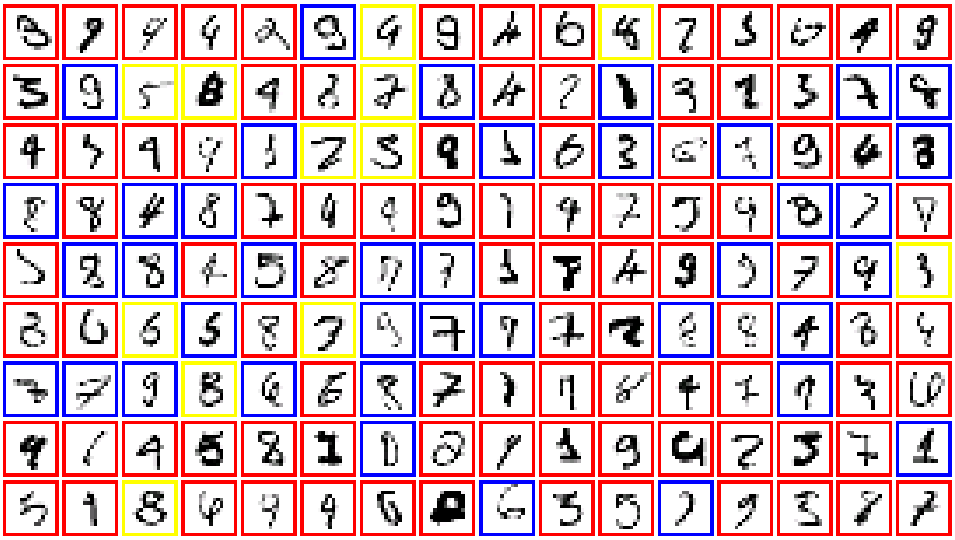
Sieci neuronowe zaimplementowano z wykorzystaniem języka programowania sprzętu VHDL. Do uruchomienia i przetestowania użyto płytę PLDA XpressFx z układem Xilinx Virtex 4-FX100.

4.1 Płyta PLDA XpressFx z układem Xilinx Virtex 4-FX100

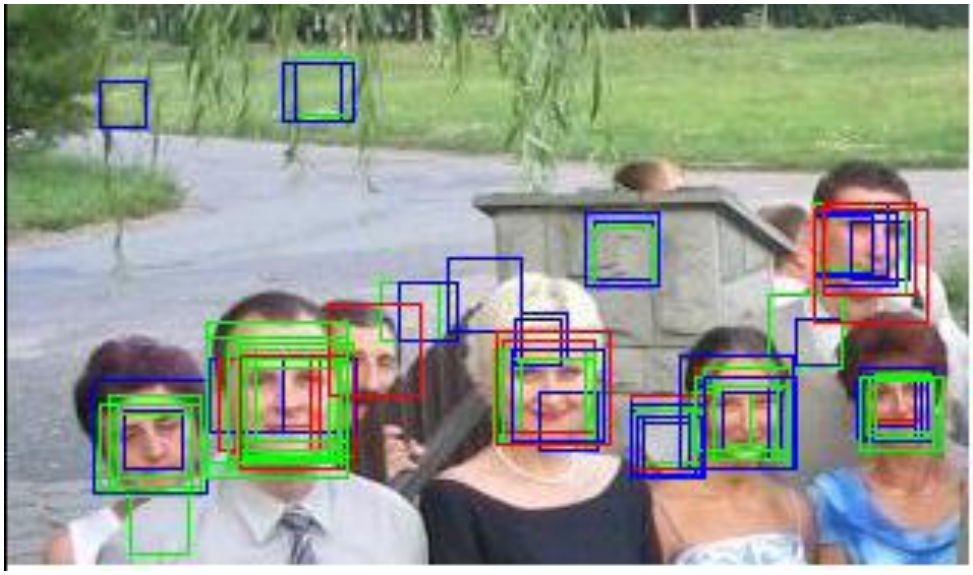
W celu otrzymania współczynników wagowych optymalnie wykorzystujących układ FPGA przedstawiono dwie metody douczania sieci neuronowej. Na podstawie eksperymentów przeprowadzonych w pracy udało się zebrać zbiór reguł służących doborowi odpowiedniej dokładności bitowej oraz zakresu wartości dla przetwarzania sieci neuronowej.

5 Koprocesor sieci neuronowych

Ponieważ układ FPGA został połączony z komputerem PC przez interfejs PCI Express, zaimplementowana w FPGA sieć neuronowa pełni rolę specjalizowanego koprocesora z zaimplementowaną siecią neuronową. Sieci neuronowe zaimplementowane w układzie FPGA użyto do akceleracji rozpoznawania ręcznie pisanych cyfr oraz systemu detekcji i rozpoznawania twarzy w obrazie. Dzięki zastosowaniu koprocesora z siecią neuronową udało się przyśpieszyć rozpoznawanie ręcznie pisanych cyfr około 100 razy w porównaniu do analogicznej logiki sieci neuronowej uruchomionej na komputerze klasy PC. System detekcji twarzy dzięki zastosowaniu sieci neuronowych uzyskał około 4-5 krotną akceleracje.



5.1 Wynik rozpoznawanie ręcznie pisanych cyfr



5.2 Wynik detekcji twarzy w obrazie

6 Wyniki

Tabela 6.1 przedstawia wyniki implementacji kilku sieci neuronowych, które z powodzeniem zostały zaimplementowane i uruchomione na układzie Virtex 4-FX100. Wszystkie implementacje były wykonane dla ograniczenia częstotliwościowego 125 MHz.

Kolumny DSP48, RAMB16 oraz SLICES przedstawiają wykorzystanie ilościowe oraz procentowe odpowiednich zasobów układu Virtex 4-FX100.

Wykorzystanie zasobów SLICES układu FPGA jest na poziomie 70%-80%. Dalsze zwiększanie pojemności sieci, a co się z tym wiąże zwiększanie wykorzystania zasobów układu FPGA nie powiodło się. Pomimo, że logika sieci neuronowej może pracować

z częstotliwością większą niż 125 MHz to po dodaniu opóźnień związanych z połączeniami poszczególnych elementów sieci neuronowej wymaganie częstotliwościowe nie było spełnione. Przykładowo dla sieci neuronowej o liczbie neuronów w kolejnej warstwie odpowiednio 196-160-10 wykorzystanie zasobów SLICES wyniosło 81%, a częstotliwość pracy spadła do 100 MHz. Dla sieci neuronowej o liczbie neuronów w kolejnej warstwie odpowiednio 196-192-10 wykorzystanie zasobów SLICES wyniosło 99%, ale częstotliwość pracy spadła do 58 MHz.

Tabela 6.1 Wyniki z implementacji sieci neuronowej dla układu Virtex 4 fx1000

Virtex 4 fx100		160		376		42176			
	DSP48	DSP48	RAMB16	RAMB16	SLICES	SLICES	Acceleration	Recognition	Recognition
Neural Network	[l.]	%	[l.]	%	[l.]	%	Ratio	PC Ratio [%]	FPGA Ratio [%]
NNFull, 4-4-4	48	30	13	3,5	9190	21,8	1,2	N/A	N/A
NNFull, 196-4-10	160	100	13	3,5	38404	91,1	70	13,0	13,5
NNFull,10-64-10	160	100	13	3,5	33962	80,5	600	N/A	N/A
NNDA 4-4-4	3	1,88	13	3,5	8124	19,3	0,4	N/A	N/A
NNDA 196-16-10	14	8,75	13	3,5	32729	77,6	200	9,0	9,8
NNNode 4-4-4	12	7,5	13	3,5	7721	18,3	0,2	N/A	N/A
NNNode 196-16-10	30	18,8	13	3,5	11678	27,7	20	9,0	9,4
NNNode 196-32-10	46	28,8	13	3,5	13391	31,8	66	4,4	4,9
NNNode 196-64-10	78	48,8	14	3,7	18777	44,5	82	3,5	3,7
NNNode 196-128-10	142	88,8	14	3,7	29179	69,2	95	2,5	2,8
NNNode 256-32-2	38	23,8	13	3,5	13252	31,4	60	5,0	5,3
NNNode 256-64-2	70	43,8	14	3,7	18467	43,8	84	4,0	4,6
NNNode 256-128-2	134	83,8	14	3,7	27970	66,3	100	2,9	3,6

7 Podsumowanie

Zaprezentowana implementacja sieci neuronowych w układach FPGA, charakteryzuje się dobrym współczynnikiem rozpoznawania, w szczególności w porównaniu z wynikami prac o podobnej tematyce. Stosunkowo skuteczne uczenie sieci związane jest ze stosunkowo dużą liczbą neuronów w sieci neuronowej. Zaprezentowane sieci neuronowe składające się z około 400 neuronów są największymi w pełni połączonymi sieciami neuronowymi, jakie zostały zaimplementowane w układach FPGA (zgodnie z dokonanym w pracy przeglądem dotychczasowych rozwiązań). Osiągnięto to przez odpowiednią optymalizację i dobór struktury sieci neuronowej. Duże znaczenie na współczynnik uczenia miał odpowiedni dobór parametrów uczenia oraz zniekształcanie wzorca uczącego. Najważniejszą zaletą implementacji sieci neuronowej w układzie FPGA jest około 100-krotna akceleracja obliczeń w porównaniu z implementacją software'ową na komputerze klasy PC.

W trakcie realizacji pracy osiągnięto wszystkie założone cele oraz została wykazana słuszność teza postawionej w rozprawie.

Zalety opracowanej w ramach pracy architektury sieci neuronowej w układzie FPGA są następujące:

- pełna parametryzacja umożliwia dopasowanie logiki sieci neuronowej do zasobów sprzętowych układu FPGA,
- mały błąd rozpoznawania ręcznie pisanych liczb na poziomie 2.8%,
- mały błąd rozpoznawania twarzy na poziomie 3.6%,
- ponad 100-krotna akceleracja obliczeń w porównaniu z rozwiązaniami programowymi,
- błąd detekcji twarzy w zakresie 15-30 %,
- 4-krotna akceleracja obliczeń systemu detekcji twarzy.

Oryginalne osiągnięcia autora pracy są następujące:

- dokonano modyfikacji uczenia sieci neuronowej metodą wstecznej propagacji błędu (*ang. backpropagation*), w tym wyrównywanie współczynników wag oraz dostrajanie współczynników wag,
- zebrano zbiór wytycznych w dotyczących doboru zakresu danych oraz dokładności reprezentacji danych liczbowych dla poszczególnych elementów sieci neuronowej,
- zminimalizowano błąd średniokwadratowy dla funkcji aktywacji aproksymowanej odcinkami,
- opracowano kod VHDL oraz implementację w układzie FPGA logiki sieci neuronowej wielowarstwowej w pełni parametryzowanej ze względu na: stopień zrównoleglenia operacji, liczbę warstw w sieci, liczbę neuronów na warstwę, dokładność obliczeń każdego z elementów sieci oraz rodzaj funkcji aktywacji,
- opracowano i uruchomiono system rozpoznawania ręcznie pisanych cyfr oraz twarzy oparty o implementację sieci neuronowej w układzie FPGA Virtex 4-FX100,
- opracowano i uruchomiono kompletny system detekcji twarzy w obrazie oparty o implementacje sieci neuronowej w układzie FPGA Virtex4 FX100.

8 Bibliografia

[1] Mariusz Kwiczala i Kazimierz Wiatr, *Implementacja sieci neuronowych w układach FPGA*. Kraków: Rozprawa doktorska, 2009.

[2] Mariusz Kwiczala i Kazimierz Wiatr, "Sprzętowa implementacja modelu neuronu w układach programowalnych FPGA," *Konferencja Metody i Systemy Komputerowe w Badaniach Naukowych i Projektowaniu Inżynierskim*, Kraków, 2003, pp. 405-410.

[3] Mariusz Kwiczala i Kazimierz Wiatr, "Sprzętowa implementacja sieci neuronowej w układach FPGA," *Konferencja Reprogramowalne Układy Cyfrowe*, Szczecin , 2005, pp. 175-182.

[4] Mariusz Kwiczala i Kazimierz Wiatr, "Sprzętowa implementacja sieci neuronowych – aproksymacja funkcji aktywacji," *Materiały V Konferencji Metody i Systemy Komputerowe*, Kraków, 2005, pp. 83-88.

[5] S. Osowski, *Sieci neuronowe w ujęciu algorytmicznym*. Warszawa: Wydawnictwa Naukowo Techniczne, 1996.

[6] Patrice Y. Simard, Dave Steinkraus, i John C. Platt, "Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis," *Microsoft Research, One Microsoft Way*, Redmond, 2003.

[7] Ryszard Tadeusiewicz, *Sieci neuronowe*. Warszawa: Akademicka Oficyna Wydawnicza, 1993.

[8] Kazimierz Wiatr, *Akceleracja obliczeń w systemach wizyjnych*. Warszawa: Wydawnictwa Naukowo-Techniczne, 2003.

[9] Kazimierz Wiatr, *Sprzętowa implementacja algorytmów przetwarzania obrazów w systemach wizyjnych czasu rzeczywistego*. Kraków: Wydawnictwo Naukowo-Dydaktyczne AGH, 2002.

[10] J. Żurada, M. Barski, i W. Jędruch, *Sztuczne sieci neuronowe*. Warszawa: Wydawnictwo naukowe PWN, 1996.