AKADEMIA GÓRNICZO-HUTNICZA

im. Stanisława Staszica w Krakowie

Wydział Elektroniki, Automatyki, Informatyki i Elektrotechniki
Katedra Automatyki
Laboratorium Biocybernetyki

1 Streszczenie

Celem naszego projektu jest zbudowanie prototypu systemu pozwalającego na detekcję twarzy na obrazie w oparciu o sieć neuronową typu CNN (Convolutional Neural Network).

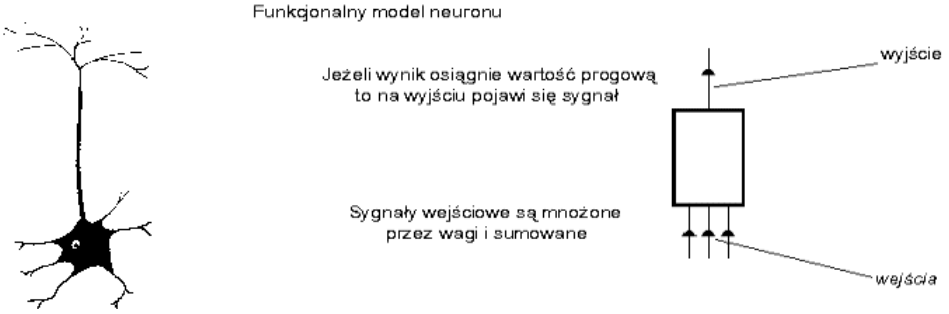
2 Wstęp

2.1 Co to jest sieć neuronowa?

Siec neuronowa jest modelem matematycznym składającym się z sieci węzłów obliczeniowych zwanych neuronami i ich połączeń. Jest to pewna technika obliczeniowo-statystyczna, należąca do dziedziny sztucznej inteligencji. Jej zadaniem jest symulacja działania ludzkiego mózgu.

2.2 Budowa sieci neuronowej

Nasz mózg jest zbudowany z ogromnej ilości neuronów które komunikują się ze sobą za pomocą impulsów elektrycznych. Z informatycznego punktu widzenia neuron stanowi układ przetwarzający pewne dane, posiadający wiele wejść oraz jedno wyjście. Oto porównanie neuronu naturalnego ze sztucznym:



Działanie sztucznego neuronu można opisać w przybliżeniu następująco:
Sygnały x_i podane na wejścia są mnożone prze odpowiednie wagi w_i oraz sumowane. Wagi mogą przyjmować dowolne wartości rzeczywiste. Równanie neuronu wygląda zatem następująco:

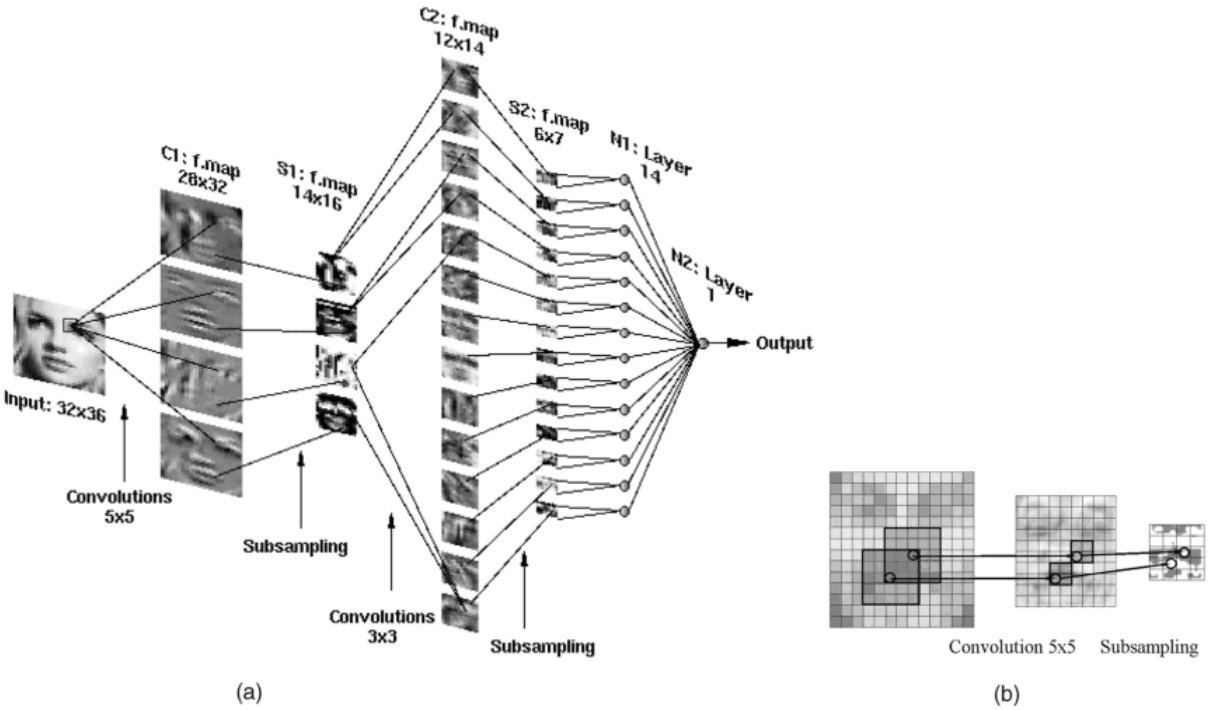
$$Y = x_1w_1 + x_2w_2 + x_3w_3 + \ldots + x_nw_n + b$$

Suma ta staje się następnie argumentem funkcji aktywacji, która zwraca wartości z zakresu (-1,1). Wynik tej funkcji pojawia się zarazem na wyjściu z neuronu.

Jak zapewne możemy zauważyć pojemność pojedynczego neuronu nie jest duza, a więc nie może on zapamiętać zbyt dużej ilości wzorców. Aby tego dokonać należy połączyć wiele pojedynczych neuronów w sieć. Dokonuje się tego tworząc kolejne warstwy. Czym więcej warstw tym sieć bardziej złożona. Ilość neuronów w poszczególnych warstwach może się różnić i zależy od konkretnego zastosowania sieci. Należy do tego dodać, że neurony w jednej warstwie nie są połączone ze sobą, natomiast neurony z dwóch sąsiednich warstw są połączone "każdy z każdym". Sygnał przechodzi do warstwy wejściowej poprzez ukrytą aż do wyjściowej, skąd jest odbierany i interpretowany.

3 Proponowane rozwiązanie

3.1 Konwolucyjne znajdowanie twarzy



Sieć ta składa się z sześciu warstw. Do warstw nie wlicza się obraz wejściowy o rozmiarach 32x36, który będzie klasyfikowany jako „twarz" lub „nie-twarz".

Warstwy od **C1** do **S2** składają się z serii obrazów na których są przeprowadzane operacje konwolucji lub subsamplingu. Obrazy te są nazywane inaczej mapami cech, ponieważ każdy z nich wydobywa z obrazu wejściowego inną cechę. Warstwa **N1** zawiera szereg niezależnych od siebie neuronów podłączonych wyjściami do **N2**, która jest zarazem zakończeniem całej sieci neuronowej. Każdy element w danej warstwie otrzymuje na wejście informacje od małego zbioru elementów z warstwy poprzedniej. Koncepcję te pokazano na rysunku **(b)**.

Postaramy się teraz opisać dokładniej poszczególne elementy tworzące przedstawioną powyżej architekturę sieci neuronowej. Różne parametry opisujące całą sieć tj. ilość warstw, ilość elementów w warstwie, rodzaj połączeń między nimi a także wielkość poszczególnych elementów została wybrana doświadczalnie przez C. Garcia i M. Delakis. Przetestowali oni szeroki zakres wielu architektur sieci i przedstawiona powyżej okazała się najbardziej skuteczna zarówno pod względem wykrywalności twarzy jak i odporności na zakłócenia obrazu.

Warstwa **C1** jest złożona z czterech map cech. Rozmiar każdej z nich to 28x32 pikseli. Każdy piksel w każdej mapie jest połączony z 5x5 sąsiedztwem z obrazu wejściowego i jego wartość odpowiada sumie ważonej tego sąsiedztwa (5x5). Dodatkowo do sumy tej jest dodawana dodatkowa zmienna trenowana zwana biasem. Podsumowując, w warstwie tej znajduje się 106(4x26) trenowanych zmiennych.

Warstwa **S1** złożona jest z czterech map cech. Każda z map powstała poprzez subsampling mapy z poprzedniej warstwy z maska 2x2, przez co jej rozmiar to połowa rozmiatu poprzedniego. Z każdych 4 sąsiednich pikseli maski **C1** jest brana średnia, która jest nastepnie mnożona przez trenowalna zmienną. Do otrzymanej wartości dodawany jest bias, po czym wynik staje się argumentem do funkcji aktywacji (tangensa hiperbolicznego). Funkcja ta używana jest po to, by uniknąć wykroczenia poza zakres zmiennych. Podsumowując, warstwa **S1** składa się z 4 map cech o rozmiarze 14x16 oraz 8(4x2) trenowalnych parametrów.

Jak mozna zauważyć na rysunku **(a)**, kolejne warstwy składają się z coraz większej ilości map cech, ale za to ich rozdzielczość się zmniejsza. Taki rodzaj systemu sieci został zaproponowany przez Hubela i Weisela i dobrze wyodrębnia on z obrazu cechy potrzebne do poźniejszego przetwarzania przez neurony.

Warstwa **C2** to konwolucyjna warstwa z 14 mapami cech. Pierwsze 8 powstało w wyniku konwolucji z maską 3x3 map z poprzedniej warstwy (każda mapa z **S1** produkuje 2 wyjścia). Pozostałe 6 to złożenia dwóch z czterech map z **S1** ($\binom{4}{2} = 6$ kombinacji).
Oto schemat przykladowego złożenia map o numerach 1 i 4:

$$Y = maska1_{3x3} * mapa_1 + maska2_{3x3} * mapa_2 + bias$$

Podsumowując, warstwa ta sklada się z 14 map cech o rozmiarze 12x14 pikseli każda oraz z 196(8x(3x3+1) + 6x(2x3x3+1)) trenowalnych zmiennych.

Warstwa **S2** to warstwa subsamplingu analogiczna do warstwy **S1**. Zmienia sie tylko ilośc i rozmiar map cech. Dostajemy zatem: 14 map o rozmiarach 6x7 oraz 28(14x2) trenowalnych zmiennych.

Ostatnie 2 warstwy składaja sie juz z neuronów. Ich zadaniem nie jest ekstrakcja cech jak miało to miejsce w poprzednich warstwach, ale ich klasyfikacja. W **N1** każdy z pojedynczych neuronów jest w pełni połączony z wszystkimi pikselami tylko jednej odpowiadającej mu mapy cech z **S2**. Natomiast w warstwie **N2** istnieje tylko jeden neuron, który łączy się z wyjściami ze wszystkich neuronów z **N1**.

Wszystkie neurony z warstw **N1** i **N2** wykonują klasyczny iloczyn skalarny pomiedzy wektorem wejściowym a wektorem wag, do czego później dodawany jest jeszcze bias. Powstała ważona suma jest nastepnie przepuszczana przez funkcję tangensa hiperbolicznego, po to,by wyjście z neuronu stanowila liczba z zakresu $(-1, 1)$.
Wyjście z ostatniej warstwy **N2** jest używana do klasyfikacji obrazu jako twarz (result>0) lub nie-twarz (result<0). Warstwy **N1** i **N2** mają zatem odpowiednio 602(14x43) i 15 trenowalnych zmiennych.