Sieci Neuronowe

Wykład 10
Wybrane zastosowania;

wykład przygotowany na podstawie.

S. Osowski, “Sieci Neuronowe w ujęciu algorytmicznym ”, Rozdz. 4, PWNT, Warszawa 1996.
W. Duch, J. Korbicz, L. Rutkowski i R. Tadeusiewicz, “Sieci Neuronowe”, Biocybernetyka i inzynieria biomedyczna 2000, t.6, ACW Exit, Warszawa 2000.

Rozpoznawanie wzorców

Przez pojęcie *rozpoznawania wzorców* rozumiemy identyfikację lub interpretację wzorca traktowanego jako obraz. Zadaniem sieci jest wyłowienie jego najważniejszych cech i zakwalifikowanie do odpowiedniej kategorii (klasy).
Można wyróżnić dwa rodzaje podejść:

→ Najpierw następuje wydobycie najważniejszych cech obrazu, a następnie sieć dokonuje na ich podstawie klasyfikacji. W wydobyciu cech obrazu są stosowane różne metody (np. momentów statystycznych)

→ Wydobywanie cech obrazu i klasyfikacja są połączone w jedno zadanie rozwiązywane przez tą samą sieć neuronową. Np. przekształcenia obrazów typu statystycznego, stanowiące fragment działania sieci neuronowej.

Przedstawimy proste podejście łączące cechy obu metod.

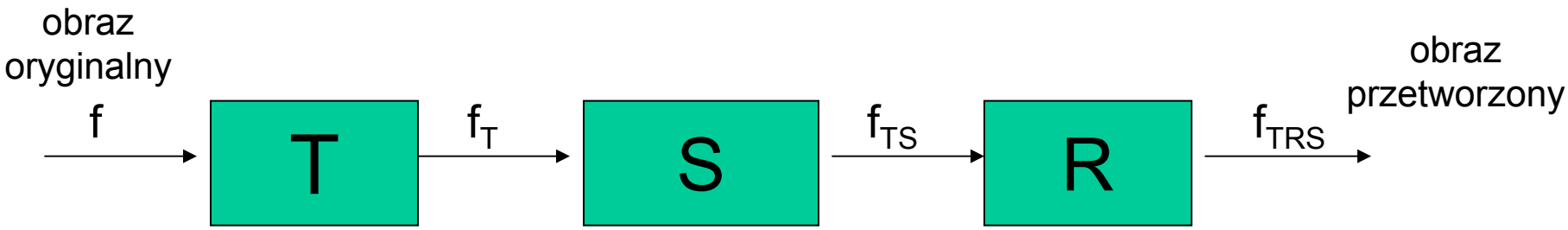
Rozpoznawanie wzorców

→ Dane dotyczące obrazu są przetwarzane przez preprocesor uniezależniający obraz od przesunięcia, rotacji i skalowania.

→ Wynik jest podawany na sieć neuronową dokonującą właściwego rozpoznania.

Główną cecha preprocesora musi być stabilność przekształcenia obrazu niezależna od poziomu szumów w obrazie oryginalnym oraz prosty i szybki w działaniu algorytm przekształcenia umożliwiający jej przeprowadzenie w czasie porównywalnym z czasem działania samego klasyfikatora neuronowego.
Jednym z takich rozwiązań jest *preprocesor o strukturze kaskadowej*.

Preprocesor o strukturze kaskadowej



Preprocesor składa się z trzech bloków:

- typu T : uniezależniający od przesunięcia wzdłuż osi x i y
- typu S : skalowanie
- typu R : rotacja

Obraz oryginalny oraz przetworzony zakodowane są w postaci pixeli.

Blok przesunięcia

Blok przesunięcia T zapewnia niezmienność względem przesunięcia na osi x i y przez określenie położenia środka ciężkości wzorca i takie jego przesunięcie, że znajdzie się ono zawsze w początku układu współrzędnych, umieszczanym zwykle w punkcie centralnym ramy obrazu. Środek ciężkości jest obliczany metodą uśredniania współrzędnych x i y wzorca.

P – liczba pixeli o przypisanej wartości binarnej 1

$$P = \Sigma \Sigma f(x_i, y_j)$$

przy czym N oznacza wymiar pixelowy ramy obrazu (przyjmuje się ramę kwadratowa), a $f(x_i, y_j)$ ma wartość binarna 0 lub 1, określającą jasność przypisana pikselowi o współrzędnych (x_i, y_j). Środek ciężkości oblicza się z zależności:

$$x_m = 1/P \Sigma \Sigma x_i f(x_i, y_j) ; \qquad y_m = 1/P \Sigma \Sigma y_i f(x_i, y_j)$$

Wzorzec wyjściowy z bloku przesunięcia określa funkcja

$$f_T(x_i, y_j) = f_T(x_i + x_m , y_j + y_m)$$

która zmienia położenie wzorca oryginalnego na płaszczyźnie umieszczając go w miejscu środka ciężkości.

Blok skalujący

Blok skalujący S to taka zmiana wymiarów wzorca, aby średnia odległość miedzy początkiem układu współrzędnych a pixelami znajdującymi się w stanie wysokim była określonym ułamkiem wymiaru ramy. Średnia odległość określa wzór:

$$r_m = 1/(\Sigma \Sigma f_T(x_i, y_j)) \; \Sigma \Sigma f_T(x_i, y_j) \, sqrt(x_i^2 + y_j^2)$$

a współczynnik skali

$$S = r_m/R$$

przy czym R jest określonym ułamkiem wymiaru ramy.
Wzorzec wyjściowy z bloku skalowania określa funkcja

$$f_{TS}(x_i, y_j) = f_T(Sx_i, Sy_j)$$

Tego typu skalowanie zapewnia ciągłość cech charakterystycznych wzorca (przy ciągłym wzorcu wejściowym f_T wzorzec wyjściowy f_{TS} jest również ciągły.

Blok rotacji

Blok rotacji R dokonuje obrotu wzorca w taki sposób, aby kierunek maksymalnej wariancji pokrywał się z osią x. Przekształcenie to wykorzystuje własność systemu, że dla danego zbioru wektorów wejściowych wektor własny stowarzyszony z największą wartością własną macierzy kowariancji wektorów wejściowych jest skierowany w kierunku maksymalnej wariancji.
Biorąc pod uwagę jedynie obrazy dwuwymiarowe, macierz kowariancji ma wymiar 2x2, dla którego wektor własny stowarzyszony z największą wartością własną, może być określony w sposób analityczny. Można doprowadzić do funkcji rzutujących postaci:

$$f_{TSR}(x_i, y_j) = f_{TS}\,(x_i \cos(\Theta) - y_j \sin(\Theta)\,,\; x_i \sin(\Theta) + y_j(\cos(\Theta))$$

gdzie $\sin(\Theta)$, $\cos(\Theta)$ odpowiadają nachyleniu wektora własnego.

Układ klasyfikatora neuronowego

→ Sygnały wyjściowe f_{TSR} preprocesora uporządkowane w postaci wektorowej składającej się z kolejnych wierszy tabeli pikselowej, stanowią sygnały wejściowe sieci neuronowej wielowarstwowej, pełniącej funkcje klasyfikatora.

→ Liczba węzłów wejściowych sieci jest równa liczbie pikseli.

→ Każdy neuron wyjściowy reprezentuje klasę, ich liczba jest również stała i równa liczbie klas.

→ Liczba warstw ukrytych i neuronów w warstwie podlega doborowi.

Klasyfikator jest trenowany metodą propagacji wstecznej przy użyciu jednego z algorytmów uczących na zbiorze danych uczących reprezentujących kolejne klasy wzorców podlegających rozpoznaniu. Biorąc pod uwagę istnienie preprocesora, wystarczy użycie jednego wzorca wejściowego dla każdej klasy.

Uklad interpretera

Na etapie rozpoznawania wzorców, biorąc pod uwagę ich zaszumienie, sygnały wyjściowe neuronów mogą przyjmować wartości ciągle z przedziału [0,1] zamiast spodziewanych wartości binarnych zero-jedynkowych z jedynką odpowiadającą rozpoznanej klasie.

→ Jednym z rozwiązań jest przyjęcie neuronu najbardziej aktywnego jako reprezentanta danej klasy.

→ Najlepszym rozwiązaniem wydaje się interpretacja dwupoziomowa:

→ sprawdza się o ile sygnał maksymalny przewyższa następny

→ jeżeli różnica jest duża za zwycięski uważa się neuron o największe aktywności

→ gdy poziomy aktywacji wszystkich neuronów są poniżej pewnego progu, interpreter ostrzega że klasyfikacja jest niepewna.

Dane literaturowe wskazują że przy bardzo prostym algorytmie przetwarzania wstępnego, tą metodą można uzyskać 90% skuteczność.

Kompresja danych

Zadaniem kompresji danych jest zmniejszenie informacji przechowywanej lub przesyłanej na odległość przy zachowaniu możliwości jej pełnego odtworzenia (dekompresji). Zastosowanie sieci neuronowej umożliwia uzyskanie nowych rozwiązań kompresji typu stratnego (z pewną utratą informacji) o dobrych właściwościach uogólniających i stosunkowo dużym współczynniku kompresji.

Sieć neuronowa do kompresji danych

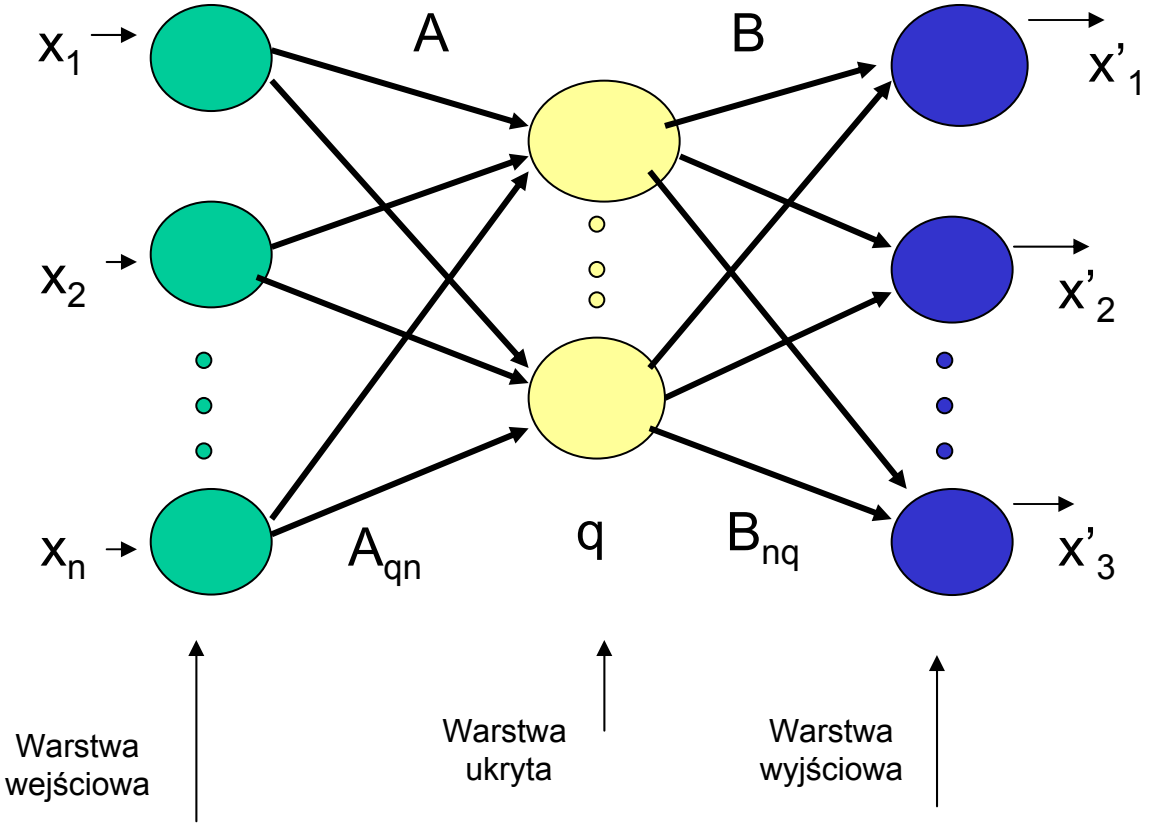
Jest to sieć dwuwarstwowa, w której liczba elementów w warstwie wyjściowej jest równa liczbie węzłów w warstwie wejściowej.

Warstwa ukryta zawiera q neuronów, przy czym $q << n$.

Warstwa wejściowa i ukryta stanowią właściwą kompresje danych, natomiast warstwa ukryta i wyjściowa realizują dekompresje.

Sieć jest typu autoasosocjacyjnego, co oznacza, że wektor uczący d jest równy wektorowi wejściowemu x, a sygnały wyjściowe sieci x'_i odpowiadają sygnałom wejściowym x_i.

Schemat sieci



Schenat sieci neuronowej jednokierunkowej do kompresji danych

Siec neuronowa wielowarstwowa

→ Kompresja dotyczy danych podzielonych na ramki (slide 16), będące ciągiem wektorów n-elementowych (n – liczba węzłów wejściowych).

→ Wobec q<<n warstwa ukryta zawiera mniejszą ilość informacji niż warstwa wejściowa, ale informacja ta prezentuje wiedzę reprezentatywną dla zbioru danych, wystarczającą do rekonstrukcji oryginalnych danych wejściowych z określoną dokładnością.

→ Warstwa ukryta reprezentuje więc składniki główne rozkładu (Principal Component Analysis – PCA), stanowiące jądro informacji.

→ Liczba tych składników jest równa liczbie neuronów q w warstwie ukrytej. Większa liczba q odpowiada zwiększonej informacji zawartej w neuronach warstwy ukrytej co z kolei umożliwia wierniejsze odtworzenie informacji wejściowej otrzymanej w wyniku dekompresji.

Siec neuronowa wielowarstwowa

Przy zastosowaniu sieci liniowej wektor h utworzony przez odpowiedzi neuronów w warstwie ukrytej oraz zdekompresowany wektor x odpowiadający sygnałom wyjściowym sieci są opisane następującymi równaniami macierzowymi

$$h = A\,x \quad \rightarrow \quad x = B\,h = B\,A\,x$$

gdzie A i B są utworzone przez wagi neuronów odpowiednio warstwy ukrytej i wyjściowej sieci.
Uczenie sieci czyli optymalny dobór wag tworzących macierz A i B wymaga aby różnica miedzy x_{ij} i x'_{ij} była dla wszystkich składowych jak najmniejsza, co prowadzi do definicji funkcji celu w postaci

$$E = \tfrac{1}{2} \Sigma\Sigma\, (x_{ij} - x'_{ij})^2$$

Nie istnieje rozwiązanie analityczne problemu (prostokątności A i B).
Uczenie neuronów: najlepsze wyniki uzyskiwano stosując liniową funkcje aktywacji.

Miary kompresji

Dane odtworzone w wyniku dekompresji są obarczone zawsze pewnym błędem. Miara tego błędu może być przyjmowana w różny sposób:

→ $MSE = d(x,x') = 1/M \Sigma (x_i - x'_i)^2$

gdzie M oznacza wymiar wektora danych x. W przypadku danych dwuwymiarowych wektor x tworzą kolejne dane dotyczące podobrazów.

Istotnym parametrem, określającym stosunek ilości informacji przypisanej obrazowi sprzed kompresji do ilości informacji odpowiadającej obrazowi skompresowanemu, jest współczynnik kompresji, K_r .

Im większy współczynnik K_r, tym większy zysk przy przechowywaniu i przesyłaniu informacji i zwykle większe zniekształcenia powstające w zdekompresowanym obrazie.

Zniekształcenie dekompresji

Zniekształcenie dekompresji mierzy się najczęściej za pośrednictwem współczynnika PSNR (Peak Signal-to-Noise-Ratio), mierzonego w decybelach i definiowanego w postaci:

$$PSNR = 10 \log((2^k-1)^2/MSE)$$

gdzie k jest liczba bitów użytych do kodowania stopni szarości obrazu. Przy 8-bitowej reprezentacji współczynnik PSNR określa wzór

$$PSNR = 10 \log (255^2/MSE)$$

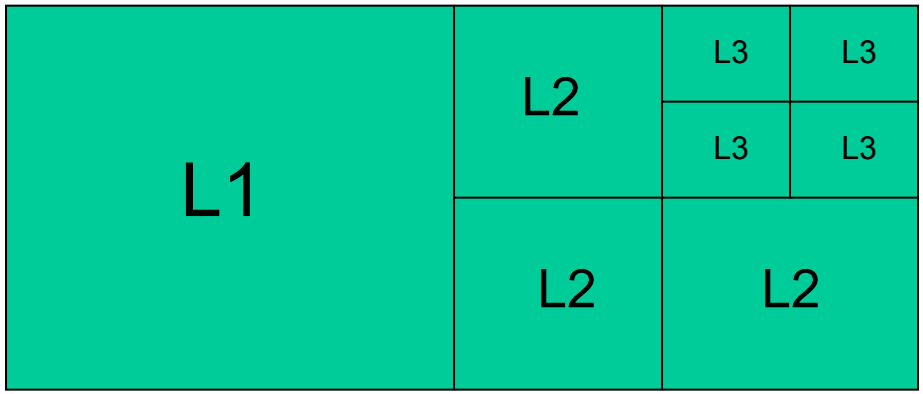
Im większa wartość współczynnika PSNR, tym lepsza jest jakość obrazu.

Hierarchiczny podział danych

Przed przystąpieniem do kompresji dane należy podzielić na ramki odpowiednich rozmiarach:

→ podział równomierny, nie uwzględnia żadnego zróżnicowania danych w poszczególnych ramkach.

→ uwzględnienie zróżnicowania, *podział hierarchiczny*. Obraz dzielony na segmenty o podobnym stopniu szarości. Segmentacja dokonywana przez regularną dekompozycję obrazu, prowadząca do struktury drzewiastej. Podział obrazu na bloki o różnych wymiarach, decyzja o kolejnym podziale jest podejmowana na podstawie pomiaru kontrastu rozumianego jako różnica miedzy najwyższym i najniższym stopniem szarości.



hierarchiczny podział obrazu

Hierarchiczny podział danych

Zastosowanie podejścia hierarchicznego w kompresji obrazów umożliwia zmniejszenie liczby wektorów uczących sieci przy zachowaniu najistotniejszych informacji zawartych w obrazie.

Zapewnienie zbliżonego do siebie kontrastu wewnątrz bloku umożliwia wydatne zmniejszenie błędu kompresji, dzięki czemu przy zadanym poziomie PSNR możliwe jest uzyskanie większych współczynników kompresji K_r .

Sieć neuronowa interpolująca

Interpolacja jest procesem polegającym na określeniu wartości funkcji w punktach pośrednich w stosunku do wartości zmierzonych.
Jej celem jest przywrócenie rzeczywistej, pełnej postaci niepełnego zbioru danych na podstawie jego fragmentów.
Przy formułowaniu matematycznych założeń przyjmuje się ciągłość funkcji oraz jej pierwszej pochodnej.

Sieć neuronowa jednokierunkowa o sigmoidalnej funkcji aktywacji może z powodzeniem spełniać funkcje układu interpolującego. Warstwa wejściowa reprezentuje dane niepełne dotyczące sygnałów zmierzonych. Warstwa wyjściowa odpowiada danym interpolowanym.

Liczba danych wyjściowych jest większa niż wejściowych, układ jest więc źle uwarunkowany i trudno jest uzyskać dobre zdolności uogólniania. Zastosowanie sieci z rozszerzeniem funkcyjnym Pao polepsza uwarunkowanie problemu interpolacyjnego i powiększa zdolności uogólniania sieci.

Modelowanie obiektów dynamicznych

W odróżnieniu od procesów statycznych, takich jak rozpoznawanie wzorca niezmiennego w czasie, w systemach dynamicznych obiekt podlegający rozpoznaniu zależy od chwilowych wartości par uczących, będących funkcją czasu.

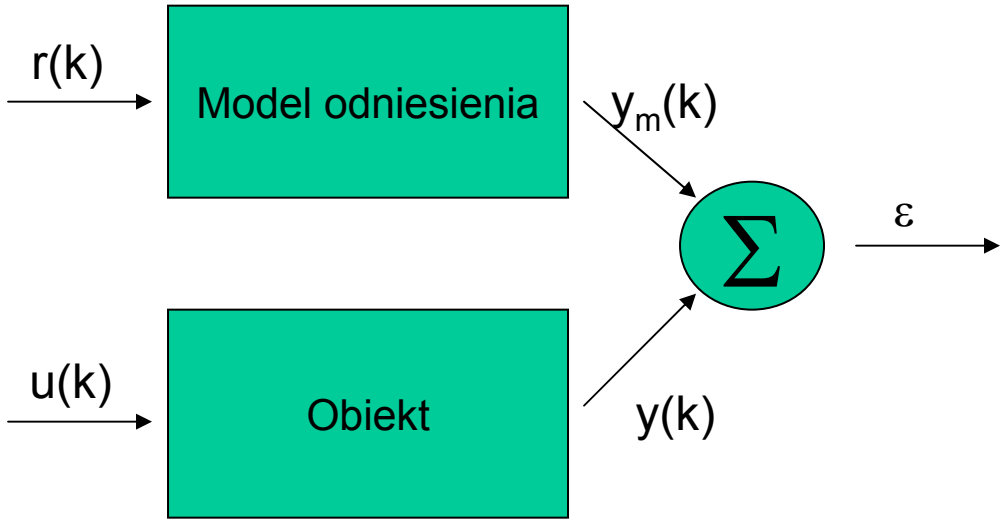
Problem identyfikacji obiektu sprowadza się do zbudowania jego modelu i określenia parametrów tego modelu w taki sposób, aby odpowiedzi obiektu y(k) i modelu y'(k) na to samo wymuszenie u(k) były sobie równe z określoną tolerancją, to znaczy

$$\| y' - y \| \leq \varepsilon$$

Sterowanie adaptacyjne znanego obiektu nieliniowego polega na doborze takiego sterowania u(k) , stanowiącego wymuszenie dla obiektu, aby odpowiedź tego obiektu y(k) śledziła i nadążała za odpowiedzią modelu odniesienia $y_m(k)$ pobudzonego sygnałem r(k).

Sterowanie adaptacyjne

Schemat układu sterowania adaptacyjnego



Wielkość $y_m(k)$ reprezentuje wielkość zadana obiektu odniesienia przy zadanym dla niego wymuszeniu r(k). Jeżeli w układzie istnieje tylko jedno wymuszenie, to zadaniem procesu adaptacyjnego jest dobór struktury i parametrów sterownika, który sygnał wejściowy r(k) przetworzy w pożądaną postać sygnału sterującego u(k), zapewniającą spełnienie warunku sterowania. Model obiektu jest zbudowany przy wykorzystaniu sieci neuronowych.

Zastosowania do predykcji

Sieci neuronowe ze względu na dobre właściwości uogólniające dobrze nadają się do rozwiązywania różnego rodzaju zadań predykcyjnych.

Szczególnie dobre wyniki uzyskuje się w systemach których działanie powtarza patern jednego z kilku możliwych wzorców, oraz jeżeli informacja ze stanu systemu w poprzednim interwale czasowym jest dobrym wskaźnikiem na to który wzorzec będzie realizowany z kolejnym przedziale czasowym.

Przykład: predykcja obciążenia sieci energetycznej

→ 4 podstawowe rodzaje wzorców obciążeń: sobota, niedziela, poniedziałek, inny dzien.,

→ każda godzina ma swoja specyfikę

→ istotną informacją jest znajomość obciążenia z poprzedniej godziny (silna korelacja)

Zestaw pytań do testu

1. Jaka jest rola preprocesora o strukturze kaskadowej przy rozpoznawaniu obrazu?
2. Jaki jest układ klasyfikatora neuronowego do rozpoznawania obrazów?
3. Jaki jest przykładowy układ sieci neuronowej do kompresji obrazów.
4. Na czym polega hierarchiczny podział danych przy kompresji obrazów?
5. Czy sieć neuronowa może pełnić rolę interpolatora?