SIECI NEURONOWE

W RÓŻNYCH DZIEDZINACH NAUKI

Bohdan Macukow i Maciej Grzenda
Instytut Matematyki, Politechnika Warszawska,
Pl. Politechniki 1, 00-661 Warszawa

1. Zamiast wstępu, czyli czym są sieci neuronowe

2. i dlaczego tak bardzo się nimi zajmujemy

Praca ta jest poświęcona sztucznym sieciom neuronowym – ciekawym i wielce obiecującym systemom przetwarzania informacji. Zainteresowanie sieciami wynika m. in. z faktu, że dostosowanie sieci do wykonania określonego zadania odbywa się zazwyczaj poprzez uczenie jej, przez użycie zestawu pobudzeń i odpowiadających im reakcji, a nie przez specjalne algorytmy, programy itp. Mówiąc o sieciach neuronowych często zamiennie używamy nazwy *neurokomputery* mając na myśli urządzenia, których budowa podobna jest do biologicznej struktury mózgu ludzkiego bądź która działa tak jak działałby mózg.

Dzisiejsze, tak szerokie i powszechne zainteresowanie sieciami neuronowymi zarówno wśród inżynierów, przedstawicieli nauk ścisłych - matematyki i fizyki oraz biologów czy neurofizjologów wynika przede wszystkim z poszukiwań nad sposobami budowy bardziej efektywnych i bardziej niezawodnych urządzeń do przetwarzania informacji a układ nerwowy jest tutaj niedościgłym wzorem. Mózg człowieka ciągle jest najpotężniejszym z istniejących obecnie urządzeń liczących do celów przetwarzania informacji w czasie rzeczywistym. Fascynacje mózgiem, jego własnościami (odpornością na uszkodzenia, równoległym przetwarzaniem itp.) już w latach 40-tych zaowocowały pracami, których fundamentalne znaczenie odczuwamy jeszcze dzisiaj.

Mózg i komputer, zastanówmy się jakie mamy tutaj podobieństwa i jakie różnice. Wyobraźmy sobie komputer, który rozwiązując pewien problem sam się uczy. Najpierw wprowadzamy do niego informacje o postawionym zadaniu, dane wejściowe problemu oraz

wybrane przykłady wraz z poprawnymi ich rozwiązaniami. Następnie komputer analizuje wprowadzone informacje i ucząc się na swoich błędach osiąga w końcu taki stan, w którym postawiony problem może być rozwiązany. W takiej działalności można zauważyć wiele podobieństwa do działania człowieka. I tutaj pojawia się pytanie: *czy potrafimy skonstruować urządzenie techniczne o podobnych właściwościach???* Badania ostatnich lat sugerują odpowiedź twierdzącą - a właśnie sieci neuronowe wydają się być drogą prowadzącą do tego celu.

Z punktu widzenia informatyki interesującym jest porównanie własności komputera z własnościami mózgu. Sieci nerwowe mogą przeprowadzać niezwykłe obliczenia i działania, aczkolwiek jest rzeczą oczywistą, że w na przykład obliczeniach arytmetycznych mózg nie jest tak dobrym urządzeniem (tak szybkim, wydajnym i dokładnym) jak komputer. Ale z drugiej strony. gdy ma się do czynienia z zadaniami takimi jak rozpoznawanie, skojarzenia czy klasyfikacja - mózg może pokonać nawet najszybszy superkomputer, pomimo że w tym procesie neurony jako jednostki przetwarzające są o wiele rzędów wielkości wolniejsze od swoich elektronicznych czy optoelektronicznych odpowiedników. Z punktu widzenia zasady działania, zarówno mózg, jak i konwencjonalne komputery realizują w zasadzie podobne funkcje - gromadzą, przetwarzają czy odzyskują informacje. Różnica nie leży więc w odmiennym działaniu lecz na odmiennych zasadach gromadzenia i przetwarzania informacji. Jeżeli mamy wykonać proste działanie arytmetyczne np. mnożenie dwu cyfr, to nie wykonujemy tego mnożenia lecz rozpoznajemy problem, a następnie przywołujemy z pamięci właściwą, skojarzoną z nim odpowiedź wynikającą z faktu, że kiedyś
w dzieciństwie uczyliśmy się tabliczki mnożenia. Jedną z najciekawszych własności, najbardziej różniącą świat sieci neuronowych od świata komputerów jest ich zdolność do tolerowania i poprawiania błędów. Mózg zdolny jest do rekonstrukcji, odtworzenia sygnału na podstawie informacji częściowej a dodatkowo obarczonej błędami.

Analizując metody przetwarzania i selekcji informacji oraz sposoby podejmowania decyzji w systemie nerwowym łatwo można dojść do wniosku, że jest on przykładem rozwiązania wielu problemów, z którymi od wielu lat boryka się informatyka, teoria przetwarzania informacji czy teoria optymalizacji.

Powróćmy jeszcze raz do genezy zainteresowania sieciami neuronowymi. Zafascynowani potęgą obliczeniową, ogromnymi szybkościami działania i możliwościami dzisiejszych komputerów często zapominamy czym naprawdę one są - jedynie doskonałymi liczydłami. Te znakomite urządzenia naprawdę są szalenie powolne i nieefektywne. Przecież do wykonania konkretnej operacji komputer wykorzystuje jedynie mikroskopijną część swoich ogromnych możliwości, używa jedynie kilku spośród swoich elementów - podczas gdy cała reszta pozostaje nieaktywna. przecież szeregowy system pracy wymusza na nim wykonywanie operacji w ustalonej

kolejności. Oczywiście staramy się temu zaradzić. Tworzymy jednostki pracujące równolegle, używamy procesorów wektorowych zwielokrotniając możliwości obliczeniowe. Jednak wszystkie te usprawnienia nikną wobec potencjalnych możliwości sieci neuronowych,—w których każdy, niezależnie pracujący neuron, jest jak gdyby niezależnym procesorem, a cała sieć zbudowana z tysięcy (lub milionów) takich procesorów może pracować w pełni równolegle i wykonywać wiele operacji równocześnie. Takie przetwarzanie pozwala sieciom niesłychanie efektywnie wykonywać złożone zadania (nawet zadania obliczeniowe), pomimo użycia bardzo powolnych elementów. Trzeba bowiem pamiętać, że typowy impuls neuronowy trwa kilka milisekund - czyli—miliony razy dłużej niż w przypadku impulsów generowanych przez krzemowe układy półprzewodnikowe.

3. Możliwości zastosowań sieci neuronowych

Omówione cechy sieci neuronowych jak także znane z literatury różnorodne modele struktur sieciowych pozwalają na scharakteryzowanie ich możliwości oraz obszarów ich potencjalnych zastosowań. Sieci nie uczą się algorytmów, lecz uczą się przez przykłady. W przeciwieństwie do konwencjonalnych komputerów są słabymi maszynami matematycznymi i słabo nadają się do typowego przetwarzania opartego o algorytmy. Bardzo dobrze natomiast nadają się do zadań związanych z rozpoznawaniem obrazów (nawet tych o niepełnej bądź zafałszowanej informacji), do zadań optymalizacyjno - decyzyjnych, do szybkiego przeszukiwania dużych baz danych.

4. Sieci neuronowe jako klasyfikatory i układy pamięciowe

Klasyfikacja jest jedną z najbardziej typowych form przetwarzania neuronowego. Jeżeli zbiór sygnałów wejściowych można podzielić na kilka klas, (nb. muszą istnieć takie cechy (atrybuty), które pozwolą na dokonanie takiego jednoznacznego podziału), to w odpowiedzi na sygnał wejściowy klasyfikator powinien podać informację o klasie, do której ten sygnał należy.

Druga grupa popularnych sieci to takie, które odtwarzają nauczony wcześniej sygnał. Takie sieci nazywamy pamięciami asocjacyjnymi. W pamięci asocjacyjnej następuje odtworzenie (odczyt) informacji zakodowanej uprzednio w pamięci. Jeżeli takiej sieci zaprezentujemy sygnał podobny do któregoś z zapamiętanych, to ma ona za zadanie te obrazy skojarzyć. Proces taki nazywamy autoasocjacją. Kojarzenie sygnałów wejściowych może także zachodzić w wariancie heteroasocjacyjnym.

W dalszej części postaram się przedstawić kilka nietypowych obszarów zastosowań sieci neuronowych. Będą to zagadnienia optymalizacji a szczególnie wykorzystanie sieci do takich zadań jak *rozwiązywanie problemu komiwojażera* czy *problemu N-hetmanów*, zastosowanie w *kompresji obrazów*, w rozwiązywaniu pewnych problemów z zakresu *algebry macierzy czy algebry liniowej* czy wreszcie w *logice.*

5. Sieci Hopfielda i Hamminga

Modele sieci Hopfielda i Hamminga są jednymi z najczęściej omawianymi, badanymi i wykorzystywanymi. Zazwyczaj obie są stosowane do rozpoznawania lub klasyfikacji obrazów, które są reprezentowane w sposób binarny. Warto także zaznaczyć, że sieć Hopfielda jest często podawana jako przykład pamięci skojarzeniowej lub jako układ do rozwiązywania zadań z zakresu optymalizacji.

Strukturę sieci Hopfielda można opisać bardzo prosto – jest to układ wielu identycznych elementów połączonych metodą *każdy z każdym*. Jest zatem najczęściej rozpatrywana jako struktura jednowarstwowa. W odróżnieniu od sieci warstwowych typu perceptron sieć Hopfielda jest siecią rekurencyjną, gdzie neurony są wielokrotnie pobudzane w jednym cyklu rozpoznawania co uzyskuje się poprzez pętle sprzężenia zwrotnego.

Wagi połączeń wyliczane są w sieci Hopfielda *a priori*, jej faza uczenia ogranicza się do wyliczenia wartości wag zgodnie zasadą uczenia Hebba

$$w_{ij} = \begin{cases} \sum_{k=1}^{M} x_i^k x_j^k & \text{dla } i \neq j \\ 0 & \text{dla } i = j \end{cases} \qquad (1)$$

Wzór 5.1

gdzie:

M jest ogólną liczbą zapamiętywanych wzorców

x_i to i-ta składowa wzorca (górny indeks określa numer wzorca)

~~$x_i \in \{-1,1\},$~~

W fazie odtwarzania na wejście sieci podany jest nieznany sygnał wejściowy i zadaniem sieci jest w procedurze rekurencyjnej „znaleźć” ten z zapisanych w jej strukturze wzorców do którego ten sygnał wejściowy jest najbardziej podobny.

Sieć Hamminga jest dwuwarstwowym klasyfikatorem o schemacie blokowym pokazanym na rys.4. Zadaniem sieci jest, podobnie jak w sieci Hopfielda, wyszukanie tego spośród zapamiętanych wzorców, który jak najbardziej podobny do sygnału wejściowego. Jako miarę podobieństwa przyjmuję się tzw. odległość Hamminga – czyli liczbę bitów rożnych w porównywanych obiektach. Tej selekcji dokonuje pierwsza warstwa klasyfikatora. Najsilniejszy sygnał wyjściowy neuronu jest wskaźnikiem najmniejszej odległości Hamminga pomiędzy sygnałem wejściowym a wzorcem klasy, którą ten neuron reprezentuje. Warstwa druga, zwana MAXNET odgrywa rolę pomocnicą. Jest to sieć rekurencyjna mająca za cel wytłumić sygnały

wyjściowe wszystkich neuronów tej warstwy oprócz twego, który otrzymał na swoim wejściu najsilniejszy sygnał wejściowy.

Wagi w pierwszej warstwie są ustalane podobnie jak w modelu Hopfielda metodą jednorazowego zapisu gdyż w praktyce są równe odpowiednim składowym zapisywanych w sieci wzorców.

Ponieważ odległość Hamminga pomiędzy wektorem wejściowym **x** oraz zapamiętanym sygnałem wzorcowym $\mathbf{s}^{(m)}$ jest równa

$$net = \begin{bmatrix} n - HD\left(x, s^{(1)}\right) \\ \ldots. \\ n - HD\left(x, s^{(p)}\right) \end{bmatrix}$$

Wzór 5.2

gdzie:

- n jest liczbą bitów w rozpatrywanych wektorach

~~, więc dodając na wejściu każdego~~ Podając na każdy element ~~elementu~~ stały sygnał wejściowy n/2, otrzymujemy łączne pobudzenie elementu W sieci MAXNET są pobudzenia pobudzające (autosprzężenie zwrotne z wagą 1) oraz hamujące, typu hamowanie oboczne, z wagą -ε, gdzie ~~0<ε<1/p~~ $0 < \varepsilon < \frac{1}{p}$ (p - ilość elementów w każdej z warstw sieci).

6. Sieci neuronowe w zastosowaniu do rozwiązywania wybranych problemów optymalizacyjnych

Dzięki swojej budowie, dzięki zdolnościom do wykonywania obliczeń równoległych oraz wynikającym stąd możliwościom przetwarzania ogromnych ilości informacji sieci neuronowe bardzo dobrze nadają się do rozwiązywania złożonych, pracochłonnych i czasochłonnych problemów optymalizacyjnych. Szybkość przetwarzania w sieciach neuronowych stwarza ogromne możliwości przyspieszenia nawet bardzo złożonych obliczeń.

Podstawowe podejście do problemów optymalizacji polega na sprowadzeniu zadania do zagadnienia minimalizacji pewnej funkcji energetycznej opisującej pewną sieć rekurencyjną traktowaną jako swoisty *układ minimalizujący*. Inne stosowane podejście, to zaprojektowanie sieci neuronowej, w której neurony wzajemnie ze sobą rywalizują dążąc do przejścia ze stanu nieaktywnego w aktywny.

Bardzo dobrym przykładem sieci neuronowej do tego rodzaju zadań jest sieć Hopfielda. Jak wiadomo, działanie jej oparte jest na samorzutnym dążeniu sieci do minimalizacji jej funkcji energii. Problem polega na odpowiednim przejściu od zadania minimalizacji funkcji celu postawionego problemu wyjściowego (z uwzględnieniem istniejących ograniczeń) do zagadnienia minimalizacji funkcji energetycznej sieci.

Wiąże się z tym konieczność rozwiązania następujących problemów:

- sposobu reprezentacji problemu przy użyciu sieci neuronowej, aby na podstawie jej stanu końcowego (wartości sygnałów wyjściowych elementów sieci) możliwe było określenie rozwiązania problemu wyjściowego
- Takiego określenie funkcji energetycznej, aby jej minimum odpowiadało optymalnemu rozwiązaniu problemu wyjściowego
- Określenia struktury, wag połączeń oraz wielkości pobudzeń zewnętrznych,
- określenia równań dynamiki poszczególnych elementów aby zapewnić zmniejszanie się wartości funkcji energetycznej,
- określenia wartości początkowych poszczególnych elementów

W typowych problemach optymalizacji kombinatorycznej funkcję energetyczną najczęściej wybiera się w postaci

$$E = \sum_i A_i(\text{miara naruszenia i-tego ograniczenia}) + B*(\text{funkcja celu})$$

Wzór 6.1

~~(4)~~

gdzie:

~~przy czym~~ A_i ~~A_i~~ i B są dodatnimi parametrami~~.~~

Poprzez minimalizację funkcji energetycznej staramy się równocześnie zminimalizować wyjściową funkcję celu oraz zmaksymalizować stopień spełnienia ograniczeń.

Takim klasycznym i dobrze znanym problemem optymalizacji kombinatorycznej jest *Problem Komiwojażera (Traveling Salesman Problem - TSP)* . Zadanie jest proste, komiwojażer ma objechać N miast. Planuje podróż w taki sposób aby każde miasto odwiedzić dokładnie jeden raz a następnie wrócić do punktu startu. Zadanie polega na minimalizacji długości przebytej trasy przy założeniu, że znamy odległości pomiędzy miastami. Problem ten posiada skończoną liczbę rozwiązań dopuszczalnych a mianowicie (N-1)!/2.

Problemy optymalizacji można podzielić na klasy według ich rozwiązywania Jeżeli istnieje algorytm, który ze wzrostem rozmiaru problemu rozwiązuje problem w czasie rosnącym tylko

wielomianowo (lub wolniej), wtedy mówimy, że jest to problem wielomianowy i należy do klasy P. Klasa P jest podklasą klasy NP podobnie jak klasa tzw. problemów NP-zupełnych. Czas potrzebny do rozwiązania problemu NP.-zupełnego wzrasta wykładniczo ze wzrostem N. Ponieważ w przypadku Problemu Komiwojażera zastosowanie pełnego przeszukiwania przy większej liczbie miast nie jest możliwe (problem ten należy właśnie do klasy NP-zupełnych) stosowane są jedynie algorytmy przybliżone, które aczkolwiek nie są pozbawione wad jednak działają w czasie wielomianowym (będącym wielomianem zmiennej N). Podstawowym problemem jest określenie odpowiedniej reprezentacji danych.

W „—__sieciowym" rozwiązaniu każde miasto jest reprezentowane za pomocą jednego wiersza zawierającego N neuronów. W każdym wierszu dokładnie jeden neuron powinien przyjmować wartość 1 a pozostałe wartość 0. Pozycja (od 1 do N), na której występuje neuron „z jedynką" odpowiada kolejności na trasie poruszania się komiwojażera.

Ogólna postać funkcji energetycznej ma postać

$$E = \frac{A}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}[(\sum_{k=1,k\neq j}^{N} v_{ik})v_{ij} + (\sum_{k=1,k\neq i}^{N} v_{kj})v_{ij}] + \frac{B}{2}\sum_{i=2}^{N}\sum_{j=1}^{i-1}[(\sum_{k=i-j+1,k\neq i}^{N} v_{k,k-i+j})v_{ij}]$$
$$+ \frac{B}{2}\sum_{i=1}^{N}\sum_{j=i}^{N}[(\sum_{k=1,k\neq i}^{N+i-j} v_{k,k-i+j})v_{ij}] + \frac{B}{2}\sum_{i=1}^{N}\sum_{j=N-i+1}^{N}[(\sum_{k=i+j-N,k\neq i}^{N} v_{k,i+j-k})v_{ij}]$$
$$+ \frac{B}{2}\sum_{i=1}^{N-1}\sum_{j=i}^{N-i}[(\sum_{k=1,k\neq i}^{i+j-1} v_{k,i+j-k})v_{ij}] + C(\sum_{i=1}^{N}\sum_{j=1}^{N} v_{ij} - N)^2 \qquad (5)$$

Wzór 6.2

gdzie:

-X, Y oznaczają miasta

natomiast i, j – etapy

, czyli v_{Xi} v_{Xi} = 1 oznacza, że miasto X zostanie odwiedzone jako i-te z kolei.

Pierwszy składnik we wzorze (5) jest równy zero wtedy ,gdy w każdym wierszu występuje tylko jedna jedynka - jest to więc rodzaj „kary" za wielokrotne odwiedzanie tych samych miast. Składnik drugi jest „karą" za równoczesny pobyt komiwojażera w dwóch różnych miejscach. Składnik trzeci równy jest zeru wtedy i tylko wtedy gdy dokładnie N neuronów jest w stanie pobudzonym (przeciwdziała zatem tendencjom minimalizacji dwóch pierwszych składników w taki sposób, aby żaden neuron nie był pobudzony). Te trzy składniki reprezentują ograniczenia problemu natomiast składnik czwarty odpowiada przebytej drodze. Jest on tak zbudowany, że sumowaniu podlegają tylko odległości d_{XY} d_{XY} między miastami kolejno odwiedzanymi. Stałe A, B, C i D dobierane są heurystycznie.

Sieć składa się z N^2 jednostek a liczba potrzebnych połączeń jest rzędu N^3. Dobre rozwiązanie zależy w znacznym stopniu od właściwego doboru stałych A, B, C i D (5). Niestety algorytm jest stosunkowo wolno zbieżny a ponadto otrzymane rozwiązanie nie jest rozwiązaniem najlepszym lecz jest rozwiązaniem optymalnym.

Bardzo podobnym problemem (należącym do tej samej klasy) jest *zagadnienie N - hetmanów* polegające na ustawieniu na szachownicy N*N, o N^2 polach, N hetmanów w taki sposób aby się wzajemnie nie atakowały. Liczba rozwiązań problemu dla kilku początkowych wartości N wynosi odpowiednio:

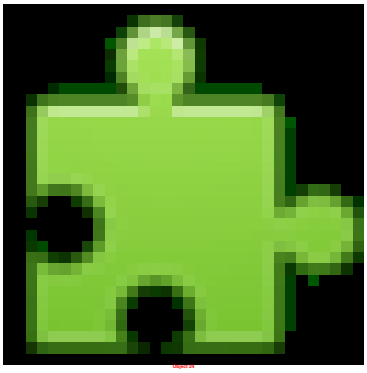
N = 1 → 1
N = 2 → 2
N = 4 → 2
N = 5 → 10
N = 8 → 92
N = 10 → 724
N = 12 → 14200 itd.

Jako reprezentację komputerową zbioru neuronów tworzących sieć rozwiązującą ten problem (o wymiarze N), zastosowano macierz kwadratową V_{N*N} ~~V[N*N];~~ Zgodnie z ideą sieci Hopfielda strukturę oraz wagi poszczególnych połączeń pomiędzy neuronami należy dobrać w taki sposób, by minima funkcji energetycznej odpowiadały rozwiązaniom problemu tzn. spełniały zbiór ograniczeń problemu. Sieć startuje z punktu „odpowiednio bliskiego" rozwiązaniu, w trakcie ewolucji będzie zmniejszała swoją energię aż do osiągnięcia stanu odpowiadającego minimum funkcji energetycznej. Zamiast zapisu V[i,j] - czyli zapisu położenia neuronu w sieci, będziemy zapisywać V_{ij} ~~V[ij]~~ jako opis odpowiadającego pola macierzy.

Sieć skonstruowano w ten sposób, że dwa różne neurony ~~(i,j)~~ (i,j) i (k,l) ~~(k,l)~~, ~~i,j,k,l∈{1,...,N}(~~ $(i,j,k,l \in 1,...,N)$, są połączone ze sobą wtedy i tylko wtedy, gdy:

- neurony znajdują się w tym samym wierszu macierzy, tj. ~~i=k~~ $i=k$, albo
- neurony znajdują się w tej samej kolumnie macierzy, tzn. ~~j=l~~ $j=l$, albo
- neurony znajdują się na tej samej przekątnej macierzy tzn. ~~i+j=k+l~~ $i+j=k+l$ albo $i-j=k-l$ ~~i-j=k-l~~.

W każdym z tych przypadków waga połączenia jest ujemna. We wszystkich pozostałych przypadkach, tzn. przy każdym innym wzajemnym położeniu neuronów (i, j) oraz (k, l) neurony te nie są połączone (czyli waga połączenia równa jest zero). Wszystkie połączenia są symetryczne. Dla tego przypadku funkcja energetyczna ma na przykład postać



(6)

Wzór 6.3

gdzie:

-A, B i C są stałymi

N_ = N+σ, σ∈<0;1). a $N = N + \sigma, \sigma \in <\{0,1\}$

Przeprowadzono 150 testów symulacyjnych dla N=8, A=B=C=100 $N = 8, A = B = C = 100$ dla wartości N– = 8.25, 8.50 i 8.75. Wybranie N=8 wynika w naturalny sposób z genezy problemu i jego praktycznej realizacji na standardowej szachownicy 64 polowej. Z dokładnej analizy problemu wynika, że dobór parametru σ ma kluczowy wpływ na jakość wyników. Najlepsze okazały się dla wartości σ z przedziału <{0,; 1}).

7. Kompresja danych

Zadanie kompresji danych polega na zmniejszeniu ilości informacji przechowywanej lub przesyłanej, przy zachowaniu możliwie jej pełnego odtworzenia (dekompresji). Stosuje się tutaj różne modele sieci i różne algorytmy ich uczenia a następnie kompresji.