Bohdan Macukow
Wydział Matematyki i Nauk Informacyjnych

SIECI NEURONOWE, HISTORIA BADAŃ I PODSTAWOWE MODELE

Do lat 40. dwudziestego wieku sieciami neuronowymi zajmowali się głównie badacze w dziedzinie neurofizjologii, analizując mechanizmy działania pojedynczych komórek układu nerwowego czy mózgu. Dopiero opracowanie pierwszego modelu neuronu oraz mechanizmu zapamiętywania informacji w sieciach biologicznych spowodowało gwałtowny wzrost zainteresowania. W latach 50. zaczęto budować pierwsze sieci neuronowe, a nawet próbowano tworzyć model całego mózgu. W niniejszej pracy przedstawiono wybrane momenty z historii tych badań oraz omówiono wybrane typowe modele sieci. Szczególną uwagę poświęcono topologii sieci jednokierunkowych. Sieci neuronowe są dzisiaj wszędzie obecne. Ich cechy, zdolność do pracy ze złożonymi, a nawet niepełnymi danymi, do analizy trendów pozwalają im rozwiązywać zadania zbyt trudne dla człowieka czy komputera.. Nowe, szybkie komputery, pozwoliły na wykorzystanie nowych metod modelowania i znacząco poszerzyły listę zagadnień, które można rozwiązać przy użyciu sieci neuronowych. Rozwój badań nad sieciami neuronowymi trwa...

1. Czym jest sieć neuronowa?

Szerokie i powszechne zainteresowanie sieciami neuronowymi wśród inżynierów, przedstawicieli nauk ścisłych – matematyków, fizyków oraz biologów i 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. Na temat sieci neuronowych (czy raczej neuropodobnych) napisano bardzo wiele prac popularnonaukowych, przeglądowych, badawczych i monografii. Były to prace pozytywnie oceniające możliwości sieci, ale było też wiele krytycznych. Przedmiotem badań były komórki lub ich elementy (dendryty, aksony, synapsy). Inna droga to badanie zespołów komórek, zasad ich funkcjonowania i współpracy. Trzecie podejście – to szeroko rozumiana percepcja czy funkcjonowanie różnorodnych funkcji życiowych. Samych definicji sieci mamy dziesiątki.
Według mnie najbardziej trafny jest opis zaproponowany przez Cichockiego i Umbehauena [5]:

„*(...) sztuczna sieć neuronowa jest układem przetwarzania informacji lub sygnałów złożonym z dużej liczby prostych elementów przetwarzających, nazywanych sztucznymi neuronami lub prościej węzłami, które są wzajemnie połączone przez bezpośrednie powiązanie nazywane wagami i które współdziałają ze sobą realizując równoległe przetwarzanie rozproszone w celu rozwiązania pożądanego zadania obliczeniowego*”.

Dodałbym jeszcze do tego opisu, że jest to bardzo uproszczony model mózgu, a topologia połączeń oraz ich parametry (tzw. wagi) są programem działania sieci, zaś reakcja sieci na zadane sygnały wejściowe jest rozwiązaniem stawianych jej zadań.

Sztuczne sieci neuronowe rozwiązują problemy w sposób inny niż konwencjonalne komputery. Algorytmiczne podejście komputerów to wykonywanie ciągu instrukcji w celu rozwiązania postawionego zadania. To ogranicza ich wykorzystanie do problemów, które rozumiemy i wiemy, w jaki sposób je rozwiązać. Sieci neuronowe działają w sposób podobny do działania mózgu. Trzeba jednak podkreślić, że metoda działania sieci neuronowych i konwencjonalnych komputerów to nie współzawodniczące, ale uzupełniające się metody, które powinny pomóc w rozwiązaniu wielu zadań, dzisiaj wydających się nie do rozwiązania.

2. Historia badań

Historię badań nas sieciami neuronowymi można podzielić na kilka okresów.

2.1. Badania podstawowe

Przed pojawieniem się historycznej pracy McCullocha i Pittsa w 1943 roku [30] prace w dziedzinie neurofizjologii skierowane były na analizę i opis mechanizmów działania mózgu czy pojedynczych komórek układu nerwowego. Prace Colgiego [6] i Ramona y Cajala [32] wykazały, że mózg składa się z neuronów, a w tej strukturze neurony nie łączą się bezpośrednio ze sobą, lecz przez połączenia synaptyczne. W 1936 roku Dale [7] zidentyfikował chemiczną naturę transmisji impulsów nerwowych. W roku 1937 Erlanger wspólnie z Gasserem [10] opisują procesy we włóknie nerwowym W latach 40. Rosenblueth wspólnie z Wienerem i Bigelowem bada analogie zachodzące pomiędzy procesami w organizmach żywych i systemach technicznych, formułując w 1943 roku podstawy współczesnej cybernetyki [35].

2.2. Pierwsze próby

Wspomniana praca McCullocha i Pittsa [30] to był prawdziwy przełom. Był to prosty matematyczny opis komórki nerwowej (neuronu), który w powiązaniu z zagadnieniem przetwarzania danych mógł modelować proste funkcje logiczne. Badania trwały i w roku 1949 Hebb [16] sformułował regułę, którą uznaje się dzisiaj powszechnie za pierwszą regułę uczenia sztucznych sieci neuronowych. Hodgkin we współpracy z Huxleyem w 1952 roku [21] stworzyli model (tzw. Model Hodgkina-Huxleya) opisujący mechanizmy jonowe leżące u podstaw procesów inicjacji i generacji potencjałów czynnościowych w neuronach. Również Eccles [9] zajmował się aspektami biofizycznymi transmisji danych w synapsach.
Większość wspomnianych tutaj badaczy za swoje prace otrzymało Nagrody Nobla.

2.3. Okres burzliwego rozwoju i wielkich nadziei

Przełom w badaniach sieci neuronowych nastąpił w drugiej połowie XX wieku. W latach 1957 i 1958 w Cornell Aeronautical Laboratory został opracowany i pomyślnie wykonany przez Rosenblatta i Wightmana [33] pierwszy neurokomputer - *Mark I Perceptron*. W 1962 r. w książce *The Principles of Neurodynamics,* Rosenblatt [34] opisuje model perceptronu. Sieć ta była zaprojektowana jako układ częściowo elektromechaniczny, częściowo elektroniczny. Nastąpił gwałtowny rozwój badań światowych nad sieciami neuronowymi tego typu. Szczególnie ciekawym rozwiązaniem jest sieć elektrochemicznych uczących się elementów Adaline, zbudowana w 1960 roku przez Widrowa i Hoffa [43] z Uniwersytetu Standforda. Sieć ta składała się z pojedynczych elementów Adaline tworząc układ Madaline.

2.4. Okres frustracji i niełaski

Rozwój badań nad sieciami neuronowymi został gwałtownie zahamowany na początku lat 70. za sprawą książki Minsky'iego i Paperta [31], która zawierała formalny dowód, że sieci jednowarstwowe mają bardzo ograniczony zakres zastosowań. Aby wykazać nieprzydatność perceptronu użyli modelu jedno-warstwowego, który jest zdolny do rozwiązania tylko problemów separowalnych liniowo, chociaż wiadomo było, że model wielowarstwowy takich ograniczeń nie ma. Był to element szeroko zakrojonej kampanii mającej na celu dyskredytację badań nad sieciami neuronowymi. Taki stan przestoju utrzymywał się przez około 15 lat.

W okresie tej „zimy" badania nad sieciami neuronowymi są prowadzone jedynie w kilku ośrodkach na świecie, Na początku lat 70. powstają pierwsze prace Kohonena [25] wykonane na Helsinki University of Technology dotyczące sieci neuronowych pamięci asocjacyjnych (skojarzeniowych). Podobne badania nad sieciami pamięci asocjacyjnych prowadził też w tym czasie Anderson w Brown University [2] i Caianiello [3] w Neapolu.

Bardzo aktywnym badaczem był w tym czasie Grossberg, który w latach 1967-1988 opublikował z zespołem 146 prac z dziedziny sieci neuronowych [np. 14]. Są to głównie prace o charakterze matematyczno-biologicznym. Fukushima i in. z NHK Laboratories w Tokio [11] proponuje szereg specjalizowanych sieci neuronowych do rozpoznawania znaków takich jak

neocognitron (to jakby pierwowzór sieci splotowej). W 1974 roku Werbos [42] opracowuje algorytm wstecznej propagacji, ale jego doniosłość została doceniona dopiero w roku 1986.

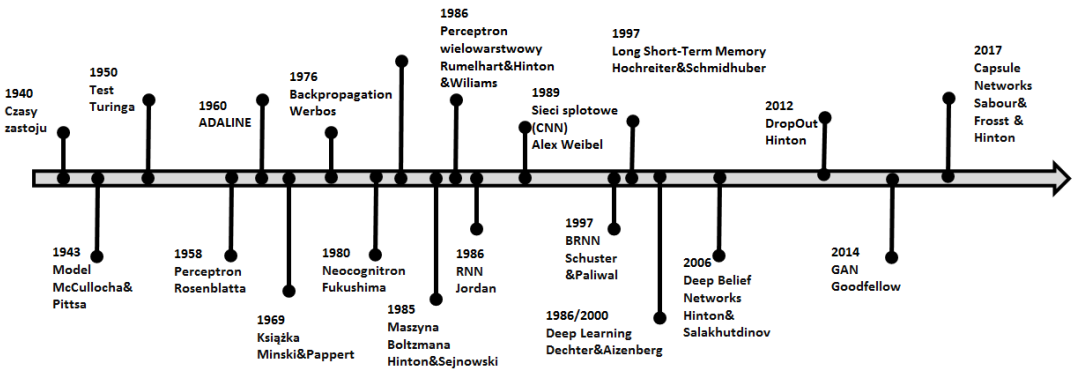
Warto zwrócić też uwagę na ówczesne dokonania polskich naukowców. Ryszard Gawroński [12] wraz zespołem w Zakładzie Bioniki Instytutu Automatyki PAN prowadzi szerokie badania nad sieciami neuronowymi. Wiele jego pomysłów powraca po latach jako np. *Sieci Komórkowe*.

2.5. Odrodzenie

W roku 1982 Hopfield publikuje przełomową pracę [22]. Opisuje rekurencyjną sieć neuronową działającą jako pamięć skojarzeniowa. Ta praca przekonuje setki naukowców , matematyków, fizyków i inżynierów, aby ponownie zająć się badaniami nad sieciami neuronowymi. W roku 1983 Hinton i Sejnowski opracowują model maszyny Boltzmana [17,20]. Algorytm wstecznej propagacji Werbosa zostaje ponownie „odkryty" i opisany w książce Rumelharta i in. [36] w roku 1986. W roku 1985 American Institute of Physics rozpoczyna organizację corocznych konferencji *Neural Networks for Computing*, a w 1987 odbywa się w San Diego pierwsza konferencja poświęcona sieciom neuronowym *IEEE International Conference on Neural Networks* i powstaje *INNS, International Neural Network Society*. Powstają specjalne wydawnictwa. W 1987 Carpenter i Grossberg [4] opisują model maszyny rozpoznającej ART1. W 1988 Kosko proponuje maszynę dwu-kierunkową (BAM) [26].

2.6. Kolejne lata

Następuje znaczne przyspieszenie badań. Przede wszystkim powstaje model Hopfielda. O ile kiedyś wiodącym modelem był prosty perceptron, teraz mamy wielowarstwowe perceptronowe sieci jednokierunkowe, a nawet sieci ze sprzężeniami. Wynalezienie sieci splotowej (konwolucyjnej - CNN) w roku 1998 rozpoczyna erę sieci bardzo dużych. Od czasów zdefiniowania *uczenia głębokiego* (*Deep learning*) wprowadzonego do uczenia maszynowego przez Dechter w roku 1986 [8], a do sieci neuronowych przez Aizenberga [1] następuje przekierunkowanie badań. Teraz badania koncentrują się na sieciach typu *Deep Belief* [18] i *ograniczonych maszynach Boltzmana* (*Restricted Boltzman Machines*) [19], w których każda warstwa komunikuje się zarówno z poprzednią, jak i następną warstwą. Opracowana zostaje *technika regularyzacji* (*DropOut*) [39] dla redukcji zjawiska nadmiernego dopasowania (*overfitting*) polegająca m.in. na rezygnacji z pewnych neuronów w różnych warstwach sieci. Goodfellow [13] proponuje *generatywne sieci współzawodniczące* (*GAN generative adversarial network*) gdzie dwie głębokie sztuczne sieci neuronowe są tak skonfigurowane, żeby ze sobą konkurowały, bez końca próbując się nawzajem pokonać. Podczas tego procesu każda z sieci staje się mocniejsza. Po raz kolejny Hinton poszerza możliwości sieci splotowej poprzez dodanie *kapsułek* (*CapsNet*) [37]. Kapsułka to zagnieżdżony zestaw warstw neuronowych. W regularnej sieci neuronowej dodaje się kolejne warstwy, natomiast w CapsNet dodaje się więcej warstw w obrębie jednej warstwy. Innymi słowy, zagnieżdża się warstwę neuronową wewnątrz innej.



Rys.1. Kamienie milowe w badaniach nad sieciami neuronowymi

Buduje się komputery wykorzystujące topologię sieci neuronowych. Pierwsze powstawały już w latach 80. W roku 1985 Mark III wyprodukowany przez TRW miał $8*10^3$ elementów, $4*10^5$ połączeń i szybkość $3*10^5$. ANZA Plus z 1988 roku to 10^6 elementów, $1{,}5*10^6$ połączeń i szybkość $6*10^6$. W roku 2016 IBM z DARPA w ramach projektu SyNAPSE stworzyły mikroprocesor TrueNorth mający 10^6 programowalnych neuronów, $256*10^6$ programowalnych synaps i $5{,}4*10^9$ tranzystorów zdolny do wykonania $4*10^{11}$ operacji synaptycznych na sekundę. Historię badań nad sieciami neuronowymi pokazuje rysunek 1.

3. Modele wybranych sieci neuronowych

Sieć neuronowa to połączone ze sobą pojedyncze elementy skonfigurowane w warstwy. Od czasów McCullocha i Pittsa powstały ogromne ilości różnorodnych modeli sieci. Różnią się funkcją, rodzajem akceptowanych danych, budową, algorytmami nauki itp. Funkcją sieci jest przetwarzanie informacji. Moc obliczeniowa sieci wynika z faktu, że wszystkie połączone elementy jednocześnie przetwarzają dane. Każdy element ma na wejściu wiele sygnałów wejściowych, których „ważność" określają wagi, ma określoną funkcję aktywacji i jedno wyjście. Działanie jest zdeterminowane przez architekturę połączeń w sieci, funkcję aktywacji i regułę uczenia. Wagi, to regulowane parametry. Ważone wejście sygnałów wejściowych to pobudzenie elementu. Dla każdego modelu sieci możemy wyróżnić trzy podstawowe fazy:

- faza uczenia sieci,
- faza testowania,
- faza wykorzystania sieci do rozwiązania nauczonego zadania.

Faza uczenia bywa bardzo różna. Dla sieci perceptronowych polega ona na wielokrotnym przedstawianiu danych uczących, a następnie na modyfikacji wag zgodnie z przyjętym algorytmem. Jest to procedura bardzo czasochłonna. W sieci Hopfielda czy w sieci komórkowej faza ucząca to na bazie postawionego zadania (danych) obliczenie wartości wag, które nie ulegają już potem żadnym zmianom.

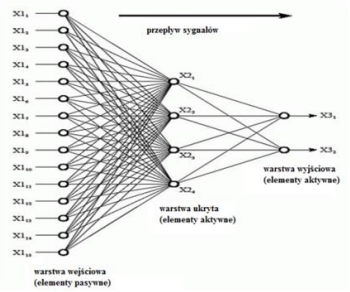
Sieć zazwyczaj definiujemy przez podanie trzech rodzajów parametrów:

a) systemu połączeń pomiędzy poszczególnymi warstwami elementów - czyli architektury,
b) reguły uczącej zmiany wag,
c) funkcji aktywacji, która przetwarza ważony sygnał wejściowy na sygnał wyjściowy.

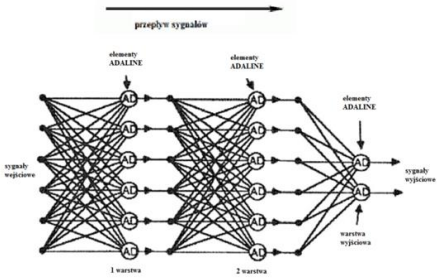
3.1 Perceptron wielowarstwowy

Jest to chyba najpopularniejsza używana dzisiaj architektura sieci (Rys.2). Perceptron to model McCullocha&Pittsa ze strategią uczenia. Uczenie perceptronu należy do grupy uczenia z nauczycielem i polega na takim doborze wag, aby sygnał wyjściowy był najbliższy wartości oczekiwanej. Strategia oparta jest na idei Rosenblatta [33], która mówi, że neuron uczy się na swoich błędach. Każdy element oblicza ważoną sumę sygnałów wejściowych (zazwyczaj z dodatkowym sygnałem polaryzacji), porównuje ją z progiem aktywacji i wykorzystując funkcję aktywacji generuje sygnał wyjściowy. Rosenblatt odkrył jedną z interesujących właściwości perceptronu: *jeżeli tylko istnieje taki układ wag, przy pomocy którego perceptron odwzorowuje w sposób poprawny zbiór wzorcowych wektorów wejściowych na odpowiadający mu zbiór oczekiwanych wartości wyjściowych, to istnieje metoda uczenia tego elementu gwarantująca zbieżność tego procesu.* Osobnym zadaniem jest określenie tego algorytmu uczącego. Dla prostego perceptronu przy zadanych wstępnie, najczęściej w sposób losowy, wartościach wag podaje się na wejście sygnał uczący i oblicza wartość sygnału wyjściowego. W wyniku porównania aktualnej wartości sygnału wyjściowego oraz wartości oczekiwanej na wyjściu dokonuje się aktualizacji wag wg zasady: *jeśli sygnał wyjściowy jest równy wartości oczekiwanej, wówczas wagi pozostają nie zmienione gdy sygnał wyjściowy jest niepoprawny dokonywane są zmiany wartości wag –*

zwiększanie gdy sygnał był zbyt mały a zwiększanie w wypadku przeciwnym. Prosty, jednowarstwowy układ, ma swoje ograniczenia (wykorzystane przez Minskiego i Papperta), dlatego z prostych elementów buduje się regularną jednokierunkową często wielowarstwową strukturę. W perceptronie wielowarstwowym wyróżnia się warstwę wejściową i wyjściową. Pozostałe, noszą nazwę warstw ukrytych. Perceptron nie zawiera połączeń pomiędzy elementami należącymi do tej samej warstwy. Połączenia pomiędzy warstwami są asymetryczne i skierowane zgodnie z ich uporządkowaniem, tzn. od warstwy wejściowej do pierwszej warstwy ukrytej, następnie od pierwszej do drugiej warstwy ukrytej, itd. aż do warstwy wyjściowej. Nie ma połączeń zwrotnych. Dla sieci wielowarstwowych najczęściej stosowanym algorytmem uczenia jest *algorytm wstecznej propagacji błędów.* Jego nazwa pochodzi stąd, że po obliczeniu sygnału wyjściowego sieci w odpowiedzi na zadany wzorzec, obliczana jest wartość gradientu funkcji błędu dla neuronów ostatniej warstwy. Następnie modyfikuje się wagi tych neuronów. Teraz błąd jest propagowany do warstwy wcześniejszej (przedostatniej). Wartości funkcji gradientu dla neuronów z tej warstwy obliczane są na podstawie gradientów obliczonych dla neuronów w warstwie następnej (czyli ostatniej). W ten sposób modyfikowane są wagi kolejnej warstwy. Takie postępowanie trwa aż do warstwy wejściowej.



Rys.2. Perceptron



Rys.3. Sieć Adaline

3.2 Adaline

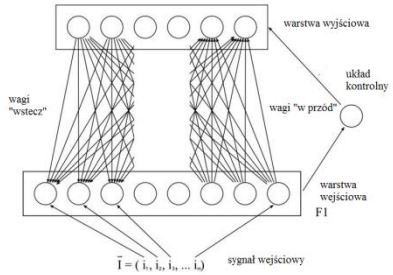
Adaptacyjny element liniowy (Rys.3) to element sieci zaproponowany przez Widrowa i Hoffa w 1960 roku [43]. Jego idea wykorzystuje klasyczny model McCullocha & Pittsa i zasady działania perceptronu. Ma sygnały wejściowe i związane w nimi wagi, sygnał polaryzacji i funkcję sumującą. O ile w klasycznym perceptronie do modyfikacji wag wykorzystuje się sygnał wyjściowy elementu (po przejściu przez funkcję aktywacji) w modelu Adaline wykorzystuje się ważony sygnał wejściowy. Z podłączania wielu elementów Adaline powstała sieć Madaline.

3.3 Sieć ART

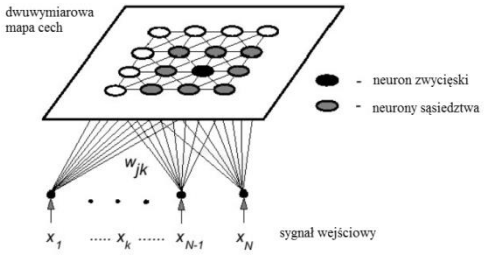
Sieć ART (*Adaptive Resonance Theory*) (Rys.4), zwana także siecią rezonansową, powstała na bazie prac Grossberga [4,14], a jej zadaniem było rozwiązanie trudnych problemów stabilności i plastyczności sieci.

Główną wadą sieci perceptronowych było ich niestabilne zachowanie. Wyrażało się ono tym, że w trakcie procesu uczenia neuron już „nauczony" na jeden sygnał często zostaje „przestawiony" i zaczyna być czuły na inny. Wymaga to powtórzenia procesu uczenia przy pomocy wzorców starych i nowych. Plastyczność, to zdolność reagowania na nowe wzorce niedające się zakwalifikować do żadnej z istniejących klas. Sieć ART jest dwuwarstwowa. Na warstwę wejściową podawany jest obraz uczący. Warstwa wyjściowa ma za zadanie wskazanie klasy, do której należy zakwalifikować obraz wejściowy. Warstwy komunikują się wzajemnie, czyli rezonują, dopasowując wagi, a proces ten powtarza się, aż do osiągnięcia maksymalnego podobieństwa. Sygnał wyjściowy z warstwy wejściowej pobudza elementy warstwy wyjściowej wyszukując „zwycięzcę" - element najmocniej pobudzony. Ten zwycięzca wysyła do warstwy

wejściowej wzorzec reprezentujący zapisaną w nim klasę. Jeżeli wzorzec jest podobny do sygnału wejściowego (jest w rezonansie) klasyfikacja jest zakończona i sygnał wejściowy zostaje zaliczony do tej klasy, a wagi połączeń ulęgają odpowiedniej modyfikacji – wzorzec zostaje „douczony". Gdy wzorzec istotnie różni się od sygnału wejściowego, układ kontrolny eliminuje z rywalizacji aktualnego zwycięzcę i sieć ponownie poszukuje zwycięzcy. Proces ten powtarza się aż do osiągnięcia wystarczającego podobieństwa, albo jeżeli sygnał wejściowy nie zostanie zidentyfikowany z żadną klasą zostaje zaakceptowany jako reprezentant nowej klasy.



Rys.4. Sieć ART



Rys.5. Dwuwymiarowa sieć Kohonena

3.4 Sieć Kohonena

Sieć Kohonena (*Self Organizing Feature Map*) (Rys.5) to przykład sieci samouczącej [25]. W odróżnieniu od uczenia, podczas samouczenia nie ma żadnego zewnętrznego źródła wiedzy (bazy danych uczących, nauczyciela itp.), dostarczającego gotowe widomości, które wystarczy tylko sobie przyswoić. Procedura samouczenia polega na grupowaniu sygnałów wejściowych, a istotą działania jest podobieństwo sygnałów. Podobne sygnały powinny pobudzać te same neurony. Jest to przykład uczenia konkurencyjnego. Sieć ma na celu utworzenie takiej struktury, która w najlepszy sposób będzie pokazywać zależności w przestrzeni sygnałów wejściowych. Sieć sama, bez „nadzoru", wykrywa istotne zależności w sygnałach wejściowych, bada ich podobieństwo, rozpoznaje cechy istotne i regularności.

Architektura sieci jest prosta – to najczęściej siatka. Każdy neuron ma ważone połączenia ze wszystkimi składowymi sygnału wejściowego, a ponadto z neuronami leżącymi w jego bezpośrednim otoczeniu. W trakcie uczenia sieci na wejście każdego elementu podawany jest *n*-wymiarowy sygnał ze zbioru wzorców uczących. We współzawodnictwie zwycięża jeden neuron, który najmocniej zareaguje na sygnał, czyli ten, którego wagi najmniej różnią się od odpowiednich składowych sygnału wzorca. Ten właśnie neuron zostaje wybrany do nauczenia (model *WTA – Winner Takes All*). Badania pokazują, że wybieranie tylko jednego neuronu nie jest najlepszym rozwiązaniem, że warto (zwłaszcza w początkowej fazie uczenia, gdy sieć jest jeszcze mocno chaotyczna) modyfikować nie tylko jeden neuron, ale również całą grupę jego sąsiadów (model *WTM – Winner Takes Most*). Następnie neuron zwycięzca i neurony sąsiedztwa podlegają adaptacji według przyjętej reguły. W konsekwencji cała przestrzeń sygnałów wejściowych zostanie podzielona na strefy wpływów poszczególnych neuronów (powstaje tzw. mapa topologiczna).

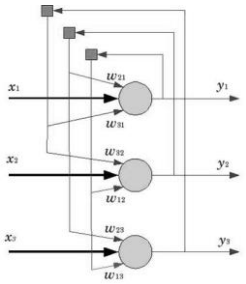
3.5 Sieć rekurencyjna Hopfielda

Sieć Hopfielda (Rys.6) wykorzystywana jest do modelowania pamięci skojarzeniowych, a także do rozwiązywania problemów optymalizacyjnych. Sieć pozwala na rozpoznanie sygnału wejściowego i odtworzenie wcześniej zapamiętanego wzorca na podstawie skojarzeń, bazując na fragmencie wzorca, wzorca podobnego lub zaszumionego [22]. Strukturę sieci Hopfielda można opisać jako układ wielu identycznych elementów połączonych każdy z każdym. Sieć Hopfielda najczęściej występuje jako struktura jednowarstwowa. W sieci Hopfielda pobudzanie zewnętrzne (sygnały x_i) jest jednorazowe, a rekurencję uzyskuje się przez pętle sprzężenia zwrotnego z wagami w_{ij}. Faza uczenia w sieci Hopfielda polega na wyliczeniu a priori (wg zasady uczenia Hebba)

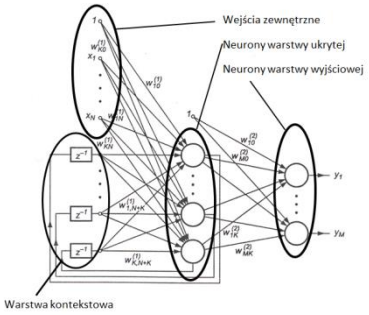
wartości wag połączeń. Pobudzona sieć, działając jako pamięć skojarzeniowa, iteracyjnie zbiega do lokalnego minimum (atraktora), ale często jest to minimum nie będące właściwym rozwiązaniem (stany odwrócone, mieszane albo obce). Procedura iteracyjna polega na minimalizacji funkcji energetycznej (funkcji błędu) sieci. Osiągnięcie na wyjściach sieci stanu ustalonego świadczy, o tym, że „skojarzyła" ona wejściowy sygnał z „podobnym" do niego zapisanym sygnałem wzorcowym. Jeśli nie udaje się osiągnąć stanu stabilnego, oznacza to, że nie potrafi ona przyporządkować sygnałowi wejściowemu żadnego z zapamiętanych wzorców. Do sieci rekurencyjnych tego typu zaliczamy również sieć Hamminga (w: [28]), czy też dwuwarstwowe uogólnienie sieci Hopfielda - zdefiniowaną przez Kosko [26] sieć BAM (*Bidirectional Associative Memory*).

3.6 Perceptronowe sieci ze sprzężeniem zwrotnym

Pod koniec lat 80. powstają nowe modele sieci oparte na sieciach jednokierunkowych typu perceptronowego wzbogacone przez dodanie sprzężeń zwrotnych. Sprzężenia te są wyprowadzane bądź z warstwy ukrytej bądź z warstwy wyjściowej i wprowadzają dodatkowe opóźnienia. Najbardziej znane z tych modeli to RMLP (*Recurrent MuliLayer Perceptron*), RTRN (*Real Time Recurrent Network*) i rekurencyjna sieć *Elmana*. Sieć Elmana (Rys.7) to sieć częściowo rekurencyjna o strukturze dwuwarstwowej. Sprzężenie zwrotne występuje między warstwą ukrytą a warstwą wejściową (jednokierunkowy przepływ sygnałów). Każdy neuron ukryty ma swego odpowiednika w warstwie kontekstowej tworzącej wspólnie z wejściami sieci sygnał wejściowy. Warstwa kontekstowa stanowi odpowiedni zestaw opóźnień jednostkowych względem wyjść neuronów warstwy ukrytej. W uczeniu sieci Elmana zazwyczaj stosuje się metodę gradientową największego spadku – często ze czynnikiem momentu.



Rys.6. Sieć Hopfielda



Rys.7. Sieć Elmana

3.7 Sieci splotowe (Convolution Neural Nets)

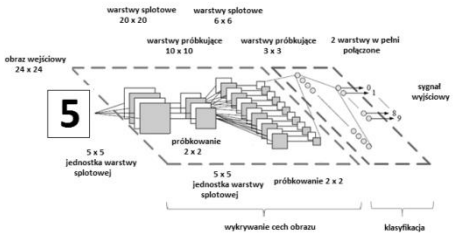
Pierwowzorem dla sieci splotowych był opracowany w roku 1980 przez Fukushimę *neocognitron* [11]. Zainspirowany pracami Hubela i Wiesla [23] stworzył wielowarstwową sieć z dwoma typami warstw: splotowymi (C) i próbkującymi (S). Sieć składała się z warstwy wejściowej (matrycy fotoreceptorów), po której następowało kaskadowe połączenie szeregu struktur modułowych, z których każda składała się z dwóch warstw komórek połączonych kaskadowo. Pierwsza warstwa każdego modułu to komórki S, które wykazują cechy podobne do prostych komórek, a druga warstwa to komórki C. Połączenia dochodzące do każdej komórki S były modyfikowalne. Sieć miała zdolność uczenia się bez nadzoru: Po kilkakrotnym zaprezentowaniu zestawu wzorców w warstwie wejściowej sieci, każdy wzór bodźca zaczął generować sygnał wyjściowy tylko z jednej z komórek C ostatniej warstwy, i odwrotnie, ta komórka C zaczęła selektywnie reagować tylko na ten wzór bodźca. Oznacza to, że żadna z komórek C ostatniej warstwy nie odpowiadała na więcej niż jeden wzór bodźca. Na odpowiedź komórek C ostatniej warstwy w ogóle nie wpływała pozycja wzorca. Nie wpływała na to niewielka zmiana kształtu ani wielkości wzoru bodźca. Sieci wielowarstwowe tego typu zaczęto nazywać

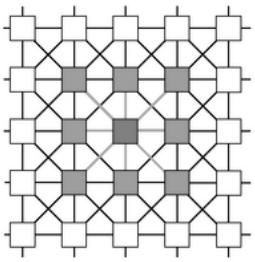
sieciami głębokimi (*deep networks*), a metody uczenia takich sieci oraz ogół zagadnień z tym związanych *uczeniem głębokim* (*deep learning*).

Lata 90. to poszukiwania efektywnych metod uczenia sieci wielowarstwowych. Dopiero w roku 2006 Hinton [18,19] pokazał, że sieci głębokie mogą być uczone w inny sposób. Zaproponowana metoda zakładała uczenie sieci warstwa po warstwie, a następnie jej nadzorowane douczenie

Rozwiązania oparte na technologii *Machine Learningu* wykorzystywane są w rosnącej liczbie branż. Jedną z najczęściej stosowanych metod *głębokiego uczenia* są tzw. *Splotowe* **(***konwolucyjne) sieci neuronowe (CNN)***.**

Pierwszy model splotowej sieci neuronowej (*LeNet-5*) zaproponował LeCun w roku 1998 [27], a potem Waibel [41]. Model ten, w wyniku szeregu zmian zaproponowanych przez badaczy, w następnych latach stał się jednym z najbardziej popularnych modeli stosowanych do przetwarzania obrazów. Na rys.8 pokazana jest przykładowa sieć splotowa w zastosowaniu do rozpoznawania cyfr.





Rys.8. Sieć splotowa Rys.9. Sieć komórkowa

Warstwy splotowe, czasami nazywane *mapą cech* (*features map*) składają się z jednostek analizujących niewielki obszar danych wejściowych, np. 5x5 pikseli. Jednostka warstwy splotowej może być postrzegana jako filtr. Dla większej liczby cech potrzebna jest większa liczba warstw splotowych. Warstwa splotowa wprowadza jednak nadmiarowość danych. Ponieważ sąsiednie jednostki częściowo pokrywają te same obszary danych wejściowych może dojść do sytuacji, że informacja o wystąpieniu cechy w danym obszarze zostaje wydobyta przez różne jednostki. W praktyce taka nadmiarowość nie wnosi wiele do modelowanego problemu, a może znacząco skomplikować jego złożoność analityczną. W celu poradzenia sobie z tym zjawiskiem wprowadza się inne rodzaj warstw - warstwy próbkujące (łączące) (*pooling layer* albo *subsampling layer*). Warstwa próbkująca ma za zadanie uogólnić informacje dotyczące danej cechy wydobyte przez jednostki warstw splotowych. Warstwy splotowe są wysoce odporne na przeuczenie, a jednocześnie obejmują niewielki obszar, dlatego nie stosuje się na nich np. metody *DropOut*. Warstwa próbkująca w praktyce uśrednia, wybiera wartość maksymalną spośród analizowanego obszaru i zmniejsza wymiarowość problemu. Pomiędzy warstwami splotowymi a warstwami próbkującymi jest warstwa ReLU o liniowej funkcji aktywacji bez nasycenia i progiem równym zero. Do tak wydobytych informacji podłączone mogą być warstwy o pełnych połączeniach znane z tradycyjnych sieci typu perceptron wielowarstwowy. Do adaptacji wag w warstwach splotowych wykorzystuje się np. zmodyfikowany algorytm propagacji wstecznej albo algorytm stochastycznego spadku gradientu, a dla warstw o pełnych połączeniach zazwyczaj algorytm propagacji wstecznej.

3.8 Sieć Komórkowa

Sieć komórkowa (*Cellural neural networks*) to układ złożony z identycznych, analogowych, nieliniowych elementów przetwarzających, zwanych komórkami, połączonych ze sobą w obrębie bliskiego sąsiedztwa i tworzących architekturę o regularnej geometrii. Wszystkie komórki przetwarzają sygnały w identyczny sposób, generując sygnał wyjściowy, którego wartość zależy od

stanu komórki wyznaczonego w drodze sumowania w czasie sygnałów sterujących komórką, przemnożonych przez odpowiednie wagi (określone macierzami sterowania i sprzężenia). Układ połączeń i rozkład wag jest dla każdej komórki w sieci taki sam. Sygnały sterujące komórką składają się z:

- sygnału wejściowego komórki,
- sygnałów wejściowych komórek należących do sąsiedztwa,
- sygnału wyjściowego komórki,
- sygnałów wyjściowych komórek należących do sąsiedztwa,
- sygnału polaryzacji.

Na rys.9 pokazany jest fragment sieci komórkowej. Komórka położona w *i*-tym wierszu i *j*-tej kolumnie oznaczona c_{ij} jest połączona bezpośrednio jedynie z komórkami należącymi do jej sąsiedztwa – w tym przypadku sąsiedztwa o promieniu równym 1. Każda komórka jest połączona z komórkami wewnątrz jej sąsiedztwa wg takiego samego schematu (sąsiedztwo komórek brzegowych jest uzupełniane sztucznie). Sieci komórkowe to przykład równoległego i rozproszonego przetwarzania obrazów. Szczególnie nadaje się do wykrywania cech lokalnych obrazów takich jak obrazy o określonej wielkości, kształcie, linii o zadanym kierunku itp.

4. Topologia, czyli innymi słowami – architektura sieci.

Jak połączyć neurony? W sieci zazwyczaj wyróżniamy elementy, które akceptują sygnały pochodzące z otocznia – tworzą one warstwę wejściową, oraz elementy, które przekazują sygnały na zewnątrz – tworzące warstwę wyjściową. Pomiędzy tymi warstwami mogą być tzw. *warstwy ukryte* (pośrednie), które pełnią niezmiernie istotną rolę w działaniu sieci. Jedną z najpopularniejszych sieci jest warstwowa sieć jednokierunkowa - sygnały wędrują od wejścia, przez warstwy ukryte do warstwy wyjściowej. Przykładem takiej sieci jest wielowarstwowy perceptron Rosenblatta czy Adaline. Poza takimi sieciami mamy także sieci rekurencyjne zawierające połączenia od warstw późniejszych do wcześniejszych. Przykładem może być sieć Kohonena czy Hopfielda. Sieci te są bardzo interesujące, aczkolwiek struktury jednokierunkowe wydają się najlepsze do rozwiązywania typowych problemów.

Czym jest warstwa sieci? Wielu autorów uważa, że warstwą jest ta część struktury, która ma zmienne wagi i dokonuje pewnych operacji (liniowych bądź nieliniowych) na sygnałach wejściowych (funkcja aktywacji). Są również tacy, którzy za warstwę uważają każdy fragment struktury złożony z elementów. Zazwyczaj elementy w pierwszej warstwie, warstwie wejściowej tylko przesyłają sygnał wejściowy do elementów warstwy następnej – bez wykonywania jakiejkolwiek operacji. Zdefiniujmy więc, że warstwa – to część struktury, w której aktywne elementy dokonują przekształcenia sygnału wejściowego.

5. Jak budować warstwową sieć jednokierunkową?

Podstawowe pytania przy budowie takich sieci:

- sieć liniowa czy nieliniowa?
- ile warstw potrzeba?
- ile ma być elementów w poszczególnych warstwach?

W sieci liniowej sygnały wejściowe po przemnożeniu przez wagi są sumowane, a powstały sygnał jest następnie przesłany do wyjścia. Czasami stosuje się tutaj charakterystykę liniową z progiem. W sieci nieliniowej, sygnał wyjściowy jest rezultatem zastosowania nieliniowej funkcji aktywacji. Przykładem może być sieć o sigmoidalnej funkcji aktywacji.

5.1 Jak wiele warstw?

Najprostsza sieć ma tylko dwie warstwy – wejściową i wyjściową (nb. sieć taka jest nazywana siecią jednowarstwową – aktywne neurony są tylko w warstwie wyjściowej). Zazwyczaj jednak pomiędzy tymi warstwami są umieszczone warstwy ukryte. Są one bardzo ważne gdyż w nich

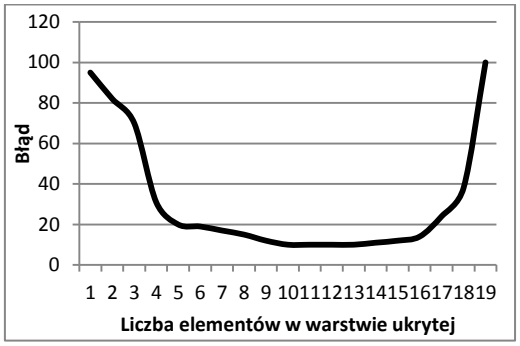
odbywa się wykrywanie cech czy kategoryzacja sygnałów. Warstwa wyjściowa to układ zebrania tych cech i generowania odpowiedzi sieci.

Sieć zbyt mała, bez warstw ukrytych, lub gdy jest zbyt mało elementów w tych warstwach, zatraca zdolność rozwiązywania problemów i nawet długie jej uczenie nie jest w stanie pomóc. Sieć zbyt duża będzie „nauczyciela" oszukiwać. Zbyt wiele warstw ukrytych, lub zbyt wiele elementów w tych warstwach, prowadzi do uproszczenia jej działania. Sieć szybko i dokładnie „nauczy się" zbioru uczącego, ale całkowicie zatraci zdolność do rozwiązywania problemów podobnych – ale nie identycznych. Badania wykazały, że do rozwiązywania większości zadań wystarcza jedna warstwa ukryta. Dodatkowe warstwy wprowadzają niestabilność gradientu i powiększają ilość fałszywych minimów. Dwie warstwy ukryte są potrzebne tylko wówczas, gdy uczenie ma na celu funkcję z punktami nieciągłości.

5.2 Ile elementów w warstwach

W warstwie wejściowej liczba elementów jest uwarunkowana rozmiarem danych. Często w warstwie wejściowej jest dodatkowy sygnał polaryzacji (*bias*). Podobnie warstwa wyjściowa jest tylko jedna, a liczba neuronów jest określona przyjętym modelem. Często stosowanym rozwiązaniem jest reguła „*jeden spośród N*", co oznacza, że odpowiedzią sieci jest niezerowy sygnał tylko jednego z elementów warstwy wyjściowej. Ponieważ jednak zazwyczaj niezerowy sygnał pojawia się na wielu elementach warstwy wyjściowej stosuje się specjalne metody preprocessingu czy specjalne progi.

Zbyt wiele neuronów w warstwie ukrytej prowadzi do powstania efektu nadmiernego dopasowania, przeuczenia (*overfitting*). Duża sieć ma duże możliwości, które nie są wykorzystane w procesie uczenia. Zwiększanie liczby elementów w warstwie ukrytej powoduje wydłużenie procedury uczącej i w rezultacie zwiększenie błędu działania sieci. Zbyt mała liczba neuronów to dla odmiany niedopasowanie, niedouczenie (*undefitting*). Często stosowaną metodą jest reguła piramidy, w której liczby elementów w kolejnych warstwach sieci maleją w kierunku od wejścia do wyjścia, często tworząc ciąg geometryczny. Inną metodą jest przyjęcie w warstwie ukrytej ilości neuronów równej średniej geometrycznej liczby elementów wejścia i wyjścia. Na rys.9 pokazana jest wielkość błędu sieci w zależności od liczby elementów w warstwie ukrytej [40].



Rys.9. Błąd działania sieci jako funkcja liczby elementów w warstwie ukrytej.

Aby określić właściwą ilość elementów w warstwie ukrytej opracowano w ostatnich 20 latach ponad 100 różnych kryteriów głównie badanych z wykorzystaniem błędu statystycznego. Dobry przegląd tych prac zawarty jest w pracy Sheela i Deepa [38].

Ciekawym rozwiązaniem może być zastosowanie technik ewolucyjnych, takich jak algorytmy genetyczne w połączeniu z algorytmem propagacji wstecznej. Zadaniem jest znalezienie struktur sieci (liczby elementów w warstwie ukrytej) zdolnej do poprawy klasyfikacji sygnałów wejściowych. Przy ustalonej liczbie elementów wejściowych i wyjściowych tworzone są populacje jako zbiór sieci o różnej liczbie elementów i różnych wagach. Funkcja oceny jest wyliczana dla

każdej sieci, operacja wyboru określa te elementy populacji które mają ocenę powyżej średniej, a operacje mutacji obejmują zmianę wag i zmianę struktury [15].

Podziękowania

Praca ma charakter przeglądowy. Do jej napisania wykorzystano część z bogatego piśmiennictwa poświęconego sieciom neuronowym. W wykazie literatury wskazano fundamentalne prace popularyzatorskie, przeglądowe [31], naukowe oraz monografie. Poza pozycjami klasycznymi w opracowaniu wykorzystano materiały z wielu opublikowanych w Internecie wykładów i innych materiałów naukowych. Część rysunków pochodzi z oryginalnych publikacji, a nawet z anonimowych źródeł w Internecie. Szczególnie wartościowy jest opis klasycznych modeli zawarty w pracach Lippmana [38] czy Husha [24].

Literatura

1. I. Aizenberg, N. Aizenberg, J. Vandewalle, *Multi-Valued and Universal Binary Neurons: Theory, Learning and Applications*. Springer Science & Business Media, 2000.
2. J. Anderson, *A simple neural network generating an interactive memory*, Mathematical Biosciences, No 14, 1972, str. 197-220.
3. E.R. Caianiello, *Reverberations and control of neural networks,* Kybernetik, No 4, 1967, str. 10-18.
4. G.A. Carpenter, S. Grossberg, *A massively parallel architecture for a self-organizing neural pattern recognition machine*, Comp. Vision, Graph. and Image Proc., No 37, 1987, str. 54-115.
5. A. Cichocki, R. Umbehauen, *Neural networks for optimization and signal processing*, 1994.
6. C. Colgi, *Sur la structure des cellules nerveuses*, Archives Italiennes de Biologie No 30, 1898, str. 60-71.
7. H.H. Dale, *Adventures in physiology*, Pergamon Press, 1953.
8. R. Dechter, *Learning while searching in constraint-satisfaction problems*, University of California, Computer Science Department, Cognitive Systems Laboratory, 1986.
9. J. Eccles, *The physiology of synapses*, Springer-Verlag, 1964.
10. J. Erlanger, H.S. Gasser, *Electrical signs of nervous activity*, Univ. of Pennsylvania Press, 1937.
11. K. Fukushima, *"Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position*, Biological Cybernetics. No 36**,** 1980, str. 193-202.
12. R. Gawroński (red.), *Bionika, system nerwowy jako układ sterowania,* PWN, 1970.
13. I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, *Generative Adversarial Networks,* Proc. 27th Int. Conf. Neural Information Processing, str. 2672-2680, 2014.
14. S. Grossberg, *Adaptive Pattern Classification and Universal Recoding*, Biol. Cyber. No 23/3, 1976, str. 121-134.
15. M. Grzenda, B. Macukow, *Genetic Algorithms to Develop the Structure of Neural Network*, Proc. Fourth Conference "Neural Networks and Their Applications", str. 593-598, 1999.
16. D. Hebb, *The Organization of behaviour*, New York, Wiley & Sons, 1949.
17. G. Hinton, T. Sejnowski, *Optimal perceptual inference*, Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, Washington DC, 1983.
18. G. Hinton, S. Osindero, Y.W. Teh, *A fast learning algorithm for deep belief nets*. Neural Computation, No 18, 2006, str. 1527-1554.
19. G. Hinton, R. Salakhutdinov, *Reducing the dimensionality of data with neural networks*, Science, No 313 (5786), 2006, str. 504-507.
20. G. Hinton, T. Sejnowski, *Learning and relearning in Boltzmann machines*, **w** [36].

21. A. Hodgkin, A. Huxley, *A quantitative description of membrane current and its application to conductionand excitation in nerve*, J.of Physiology, No 117, 1952, str. 500-544.
22. J. J. Hopfield, *Neural networks and physical systems with emergent collective computational abilities*, Proc. Natl. Acad. Sci. USA, No 79, 1982, str. 2554-2558.
23. D. H. Hubel, T.N. Wiesel, *Receptive fields of single neurones in the cat's striate cortex*, J. Physiol. No 148/3, 1959, str. 574–591.
24. D. R. Hush, B. H. Horne, *Progress in Supervised Neural Networks*, IEEE Sign Proc.Mag., Jan. 1993.
25. T. Kohonen, *Self-organized formation of topologically correct feature maps*, Biol. Cyber. No 43, 1982, str. 59-69.
26. B. Kosko, *Bidirectional associative memories*, IEEE Transactions on Systems, Man, and Cybernetics, No 18/1, 1988, str. 49-60.
27. Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, *Gradient-Based Learning Applied to Document Recognition*, Proceedings of the IEEE, No 86/11, 1998, str. 2278-2324.
28. R. Lippmann, *An Introduction to Computing with Neural Nets*, IEEE ASSP Magasin, str. 4-22, April 1987.
29. B. Macukow, *Neural Networks - State of Art, Brief History, Basic Models and Architecture*, Proc. 15th IFIP TC8 International Conference, CISIM 2016, LNCS No 9842, 2016, str. 3-16.
30. W.S. McCulloch, W. Pitts, *A logical calculus of the ideas immanent in nervous activity*, Bull. Math. Bioph. No 5, 1943, str. 115-133.
31. M.L. Minsky, S. Papert, *Perceptrons: An introduction to computational geometry*, Cambridge, MIT Press, 1969.
32. S. Ramon Y. Cajal, *Les nouvelles idées sur la fine anatomie des centres nerveux,* C.Reinwald &C^{ie}, Paris, 1894.
33. F. Rosenblatt, *The Perceptron: A probabilistic model for information storage and organization in the brain*, Psychological Review, No 65/6, 1958, str. 386–408.
34. F. Rosenblatt, *Principles of neurodynamics*, Washington, Spartan Books, 1962.
35. A. Rosenblueth, N. Wiener, J. Bigelow, *Behavior, purpose and teleology,* Philosophy of Science, No 10, 1943, str. 18–24.
36. D.E. Rumelhart, G.E. Hinton, R.J. Williams, *Learning internal representations by error propagation*, w: Parallel distributed processing: explorations in the microstructure of cognition, Vol. 1, 1986 str. 318-362, MIT Press, Cambridge.
37. S. Sabour, N. Frosst, G.E. Hinton, *Dynamic Routing Between Capsules*, Proc. 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, 2017, str. 3859–3869.
38. K.G. Sheela, S.N. Deepa, *Review on Methods to Fix Number of Hidden Neurons in Neural Networks*, Math. Probl. Eng., Vol. 2013, India, 2013.
39. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, *Improving neural networks by preventing co-adaptation of feature detectors*, Journal of Machine Learning Research No 15, 2014, str. 1929-1958.
40. R. Tadeusiewicz, *Odkrywanie właściwości sieci neuronowych,* PAU Kraków, 2007.
41. A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K.J. Lang, *Phoneme Recognition Using Time-Delay Neural Networks*, IEEE Transactions on Acoustics, Speech, and Signal Processing, No 37/3, 1989, str. 328. – 339.
42. P. Werbos, *Beyond regression: new tools for prediction and analysis in the behavioral sciences*, Ph.D. thesis, Harvard University, Cambridge, 1974.
43. B. Widrow, M.E. Hoff, *Adaptive switching circuits*, IRE WESCON Conv. Record, NY, Vol. 4, 1960, str. 96-104.