Metody głębokiego uczenia w rozpoznawaniu obrazów

Jacek Witkowski
Wydział Elektroniki i Technik Informacyjnych, Politechnika Warszawska,
ul. Nowowiejska 15/19, 00-665 Warszawa
j.witkowski@stud.elka.pw.edu.pl

Streszczenie. *W artykule zwrócono uwagę na dynamiczny rozwój sztucznej inteligencji w ostatnich czasach. Następnie objaśniono różnice pomiędzy uczeniem nadzorowanym i nienadzorowanym oraz zaprezentowano przykładowe zastosowania uczenia maszynowego we współczesnym świecie. Omówiono również temat sztucznych sieci neuronowych: ich budowę oraz zasadę działania. Następnie wyjaśniono czym jest uczenie głębokie. Kolejno, zaprezentowano szczególny przypadek sieci neuronowej, jakim jest sieć splotowa. Wyjaśniono, w jaki sposób jest ona wykorzystywana do rozpoznawania obiektów znajdujących się na obrazach. Zaznaczono również, że zagadnienie to jest tematem pracy magisterskiej autora artykułu.*

1. Motywacja. Ostatnimi czasy sztuczna inteligencja jest wykorzystywana w coraz większej liczbie obszarów. Zaczyna ona być obecna nawet w urządzeniach codziennego użytku. Szybki rozwój uczenia maszynowego, czyli dziedziny zajmującej się badaniem sztucznej inteligencji, stał się możliwy dzięki osiągnięciu znacznych mocy obliczeniowych komputerów. Pozwoliło to na stosowanie zaawansowanych algorytmów, które kiedyś były jedynie przedmiotem rozważań teoretycznych.

Biorąc pod uwagę to, że sztuczna inteligencja może być stosowana w prawie każdej dziedzinie nauki, warto zgłębić wiedzę na temat najnowszych osiągnięć związanych z uczeniem maszynowym. W przeciągu ostatnich 5 lat wyjątkowo szybko rozwijały się algorytmy służące do rozpoznawania obrazów. Prawdopodobnie było to spowodowane tym, że jest to obszar, w którym można znaleźć wiele zastosowań dla sztucznej inteligencji.

Jednym z mechanizmów najczęściej wykorzystywanych do rozpoznawania obrazów oraz materiałów wideo jest tzw. sieć splotowa, która zostanie dokładniej omówiona w dalszej części artykułu.

2. Uczenie nadzorowane i nienadzorowane. Mechanizmy sztucznej inteligencji można podzielić na:

- uczenie nadzorowane,
- uczenie nienadzorowane.

Uczenie nadzorowane to proces, w którym wraz z danymi wejściowymi sieci, dostarczamy do uczonego mechanizmu również wynik spodziewany na jego wyjściu. Wówczas celem uczenia jest minimalizacja różnicy pomiędzy danymi wygenerowanymi przez mechanizm, a danymi spodziewanymi.

W **uczeniu nienadzorowanym** w procesie uczenia dostarczane są jedynie dane wejściowe (bez spodziewanego wyjścia, czyli tzw. etykiet). Wówczas celem uczenia jest modelowanie danych wejściowych (a dokładniej: rozkładu prawdopodobieństwa danych wejściowych). Przykładowo: przy uczeniu sieci neuronowej w sposób nienadzorowany, jeśli będzie ona otrzymywała obrazki ludzkich twarzy, to będzie w stanie rozpoznawać często występujące zależności pomiędzy pikselami (np.: krawędzie, nos, oczy, usta, jak również twarze).

Oba rodzaje uczenia mogą być łączone. W ten sposób działa również mózg człowieka. By zobrazować omawiane podejście, warto posłużyć się przykładem. Wyobraźmy sobie, że ktoś chce nas nauczyć rozpoznawać różne rodzaje jabłek. Najprostszym po-

dejściem byłoby pokazywanie nam wielu jabłek wraz z opisem wskazującym jakiego rodzaju jest każde z nich. Byłoby to jednak bardzo czasochłonne, gdyż zaprezentowanie każdego jabłka wymagałoby przygotowania dla niego odpowiedniego opisu. Łatwiejszym sposobem byłoby zamknięcie nas w pokoju z wieloma nieopisanymi jabłkami. Wówczas zaczęlibyśmy oglądać każde z nich i po pewnym czasie zauważylibyśmy różne cechy powtarzające się na niektórych jabłkach (np. podłużny kształt, czerwony kolor, zielony kolor, zielone plamki itp.). Po wyjściu z pokoju potrafilibyśmy identyfikować cechy występujące w różnych grupach jabłek. Wiedza ta pozwoliłaby nam na nauczenie się rozpoznawania różnych gatunków na podstawie mniejszej liczby opisanych przykładów niż przy podejściu naiwnym, gdzie od początku stosowane jest uczenie nadzorowane.

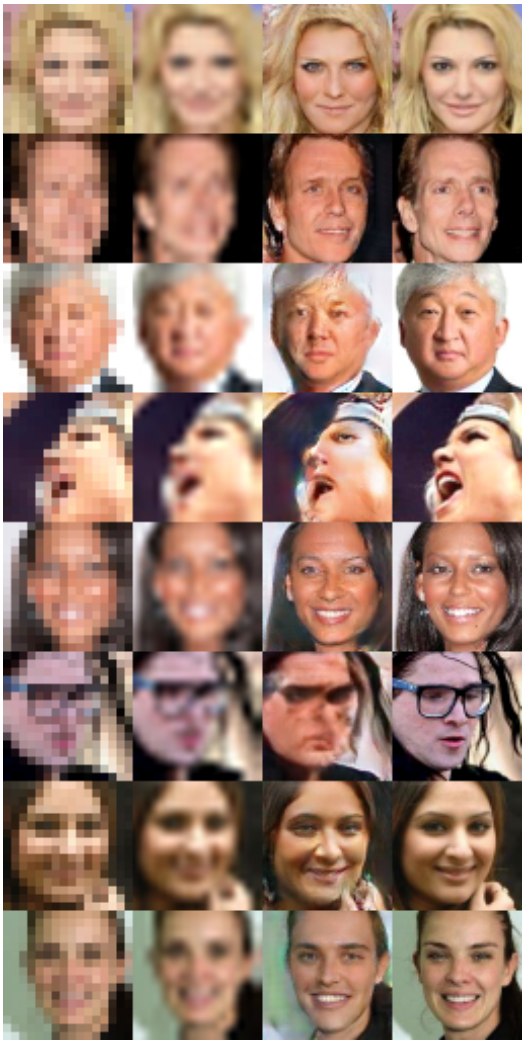
3. Przykłady zastosowań.

3.1. Ocena atrakcyjności zdjęć. By lepiej zrozumieć, jakie możliwości otwiera przed nami obecny rozwój uczenia maszynowego, warto poznać zastosowania, jakie znajduje ta dziedzina już teraz. Jednym z nich jest zastosowanie Głębokiej Splotowej Sieci Neuronowej (*ang. Convolutional Neural Network, CNN*) do oceniania atrakcyjności zdjęć typu „selfie" [1]. W omawianym projekcie autor:

1. Zgromadził zdjęcia, które oznaczone były hasz-tagiem „selfie" (znalazł około 5 milionów takich zdjęć).
2. Przy użyciu mechanizmu rozpoznającego twarze na obrazkach wybrał tylko te zdjęcia, na których obecna była przynajmniej jedna twarz (zostało około 2 milionów zdjęć).
3. Posortował fotografie względem liczby użytkowników obserwujących autora danego zdjęcia.
4. Podzielił posortowaną listę fotografii na grupy po 100 zdjęć (każda grupa dzięki temu zawierała zdjęcia o podobnej liczbie obserwujących danego użytkownika).
5. W obrębie każdej grupy, utworzył ranking zdjęć na podstawie ich liczby polubień. Górna połowa zdjęć była uznawana za zdjęcia dobre, a dolna - za złe.
6. Trenował sieć neuronową milionem zdjęć dobrych oraz milionem zdjęć złych. Dzięki temu mechanizm nauczył się odróżniać oba typy zdjęć.

Tak utworzona splotowa sieć neuronowa była w stanie ocenić prawdopodobieństwo, że dane zdjęcie jest atrakcyjne (im wynik na wyjściu sieci był wyższy, tym zdjęcie było lepiej oceniane).

3.2. Image Super Resolution. Innym przykładem zastosowania uczenia maszynowego jest projekt Image Super Resolution [2]. W swojej aplikacji autor użył sztucznej inteligencji w celu czterokrotnego powiększania obrazków: z rozmiaru 16x16 pikseli do 64x64 piksele. Porównanie wyników działania sieci neuronowej oraz interpolacji bikubicznej (standardowej metody stosowanej do powiększania obrazów w programach graficznych) zostały przedstawione na rysunku 1. Architektura zastosowana w projekcie to DCGAN (*ang. Deep Convolutional Generative Adversarial Network*, Głęboka Generatywna Antagonistyczna Sieć Splotowa [3]). Wykorzystuje ona uczenie nienadzorowane. W uproszczeniu: autor mechanizmu w procesie uczenia używał obrazków przedstawiających twarze. Sieć nauczyła się wówczas jakie są najczęstsze zależności pomiędzy pikselami występujące w tych obrazkach, a więc mogła również generować twarze. Etap ten był uczeniem nienadzorowanym. Następnie autor zastosował kolejny uczenie nadzorowane, w którym jako funkcję błędu wykorzystał odległość L1 pomiędzy wynikiem powiększania obrazka przez sieć, a oryginalnym obrazkiem o rozmiarze 64x64 piksele. Odległość L1 (inaczej odległość Manhattan) jest zdefiniowana jako suma różnic odpowiadających sobie współrzędnych dwóch wektorów, tj. $\sum_{i=0}^{n}(x_i - y_i)$, gdzie n to liczba współrzędnych (w tym wypadku: liczba pikseli pomnożona przez liczbę kanałów, która dla obrazu kolorowego jest równa 3), a x i y to porównywane obrazki.



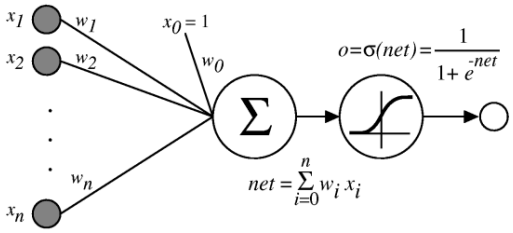
Rysunek 1. Powiększanie obrazków (pierwsza kolumna zawiera powiększany obraz, druga - obraz powiększony poprzez zastosowanie interpolacji bikubicznej, trzecia - obraz powiększony przez sieć splotową, czwarta - oryginalny obraz 64x64 piksele).

3.3. AlphaGO. Do bardziej zaawansowanych sztucznych inteligencji należy projekt „AlphaGo" [4] stworzony przez firmę Google DeepMind. Owa sztuczna inteligencja nauczyła się grać w starochińską grę planszową - Go. Mechanizm wyróżnia to, że może być wykorzystany do różnych celów (np. może nauczyć się grać w inne gry) i nie był tworzony pod z góry określone zastosowanie, w przeciwieństwie do mechanizmów takich jak: IBM Watson czy IBM Deep Blue. Silnik stworzony przez firmę Google DeepMind jako dane wejściowe przyjmuje obraz w postaci mapy bitowej, a następnie do jego przetwarzania wykorzystuje głębokie splotowe sieci neuronowe (opisane w dalszej części artykułu) oraz Q-learning (szczególny rodzaj uczenia ze wzmocnieniem).

Program AlphaGo jest przełomowy również pod innym względem: jest pierwszą sztuczną inteligencją, która pokonała profesjonalnego gracza w Go [5] na pełnowymiarowej planszy (19x19) bez zastosowania tzw. handicapu. Pojedynek odbył się w październiku 2015 roku. Niedługo potem, w marcu 2016 roku, AlphaGo zdołał pokonać, 18-krotnego i również ówczesnego mistrza świata w grę Go [6].

Tym samym ostatnia z popularnych gier planszowych straciła mistrza ludzkiego na rzecz programu komputerowego (podobnie jak wcześniej warcaby [7] czy szachy [8]).

4. Sieci neuronowe. Jednym z mechanizmów, który jest najczęściej wykorzystywany do implementacji sztucznej inteligencji, jest sztuczna sieć neuronowa. Jest ona inspirowana sieciami neuronowymi występującymi w biologii (np. w ludzkim organizmie, w szczególności: w mózgu). Na strukturę sieci składają się pojedyncze połączone ze sobą elementy, zwane neuronami. Budowa neuronu została przedstawiona na rysunku 2.



Rysunek 2. Neuron Hebba.

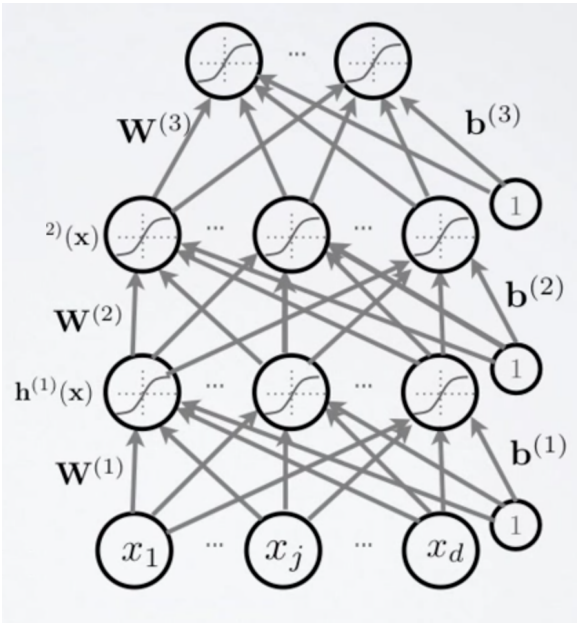
Neurony najczęściej łączone są w tzw. sieci warstwowe. Przykładowa sieć warstwowa została zaprezentowana na rysunku 3. Jeśli liczba warstw występujących w danej sieci jest dostatecznie duża (zazwyczaj: około 10 warstw lub więcej), sieć określana jest jako głęboka.

Algorytm, który pozwala otrzymać wartość wyjściową dla neuronu, przedstawia się następująco:

1. Pobierz dane wejściowe, które są wektorem: $\overline{X} = [x_1, x_2, ..., x_n]$.
2. Pomnóż każdą współrzędną wektora wejściowego przez odpowiadającą

mu wagę (określaną w procesie uczenia sieci) i zsumuj ze sobą otrzymane wartości: $net = \sum_{i=0}^{n} w_i x_i$.

3. Do otrzymanego wyniku dodaj czynnik b zwany przesunięciem (*ang. bias*): $net := net + b$.
4. Otrzymaną sumę poddaj działaniu funkcji σ zwanej funkcją aktywacji: $o = \sigma(net)$. Wynik tej operacji jest wartością wyjściową neuronu.



Rysunek 3. Warstwowa sieć neuronowa typu Feed-Forward. Źródło: [9]

5. Splotowe sieci neuronowe. Szczególnym przypadkiem sieci neuronowych są sieci splotowe. Wykorzystują one operację splotu w postaci dyskretnej:

$$(f * g)[n] = \sum_{m=-\infty}^{+\infty} f[m]g[n-m] = \sum_{m=-\infty}^{+\infty} f[n-m]g[m]$$

Dzięki stosowaniu różnych masek (funkcji splatanych z przetwarzanym sygnałem, np. obrazem) w filtrach splotowych można uzyskiwać różne efekty. Istnieje wiele zdefiniowanych funkcji tego typu, które służą m.in. do:

- redukcji szumów,
- wyostrzania,
- wykrywania krawędzi.

Sieci splotowe wykorzystują to, że odpowiednio dobierając wartości maski, można uzyskiwać bardzo różne efekty. Jednak maska zamiast przyjmować z góry zadane wartości (tak jak w standardowych filtrach splotowych), otrzymuje je w procesie uczenia.

Standardowym zadaniem, do którego wykorzystywane są splotowe sieci neuronowe, jest identyfikacja obiektu obecnego na zdjęciu. Algorytm przetwarzania obrazu dzieli się na etapy:

1. Splot: zastosowanie n różnych filtrów splotowych, wynikiem czego jest n obrazków zwanych mapami cech.
2. Normalizacja (krok opcjonalny, mający na celu taką zmianę danych, która poprawi jakość uczenia sieci). Krok ten wyjaśniono w dalszej części artykułu).
3. Próbkowanie (zmniejszenie rozmiarów powstałych map cech, m.in. w celu redukcji rozmiaru przetwarzanych danych).
4. Powtórzenie kroków 1-3 wiele razy (liczba powtórzeń jest zależna od liczby warstw w sieci).
5. Przekazanie powstałych map cech do warstwy neuronów w pełni połączonej (każdy piksel trafia do wszystkich neuronów tej warstwy sieci).
6. Zastosowanie kolejnych warstw w pełni połączonych (zazwyczaj w sieci stosuje się jedną lub dwie takie warstwy). Ostatnia z tych warstw ma za zadanie przedstawić na swoim wyjściu prawdopodobieństwa występowania różnych etykiet (np. kot: 50%, pies: 20%, samochód: 5%, ...)

Drugi krok, czyli normalizacja, jest krokiem, w którym zawierają się różne operacje mające na celu taką zmianę danych, by proces uczenia uległ poprawie (np. szybsze uczenie sieci, lepsza jakość klasyfikacji w sieci). Jest to zagadnienie bardzo obszerne i nie jest omawiane w tym artykule.

Splotowe sieci neuronowe szczególnie często są wykorzystywane do przetwarzania obrazów i filmów. Jest to spowodowane tym, że mogą one operować na danych wielowymiarowych

(tj. obrazach czy filmach), a przy tym potrafią znajdywać lokalne cechy występujące w sygnale (np. człowiek może pojawiać się w różnych miejscach na zdjęciu, a sieć zawsze będzie w stanie określić, że fotografia przedstawia człowieka).

6. Podsumowanie. Z powodu gwałtownego wzrostu mocy obliczeniowej w ciągu ostatnich 10 lat znacząco zwiększyły się możliwości mechanizmów wykorzystujących sztuczną inteligencję. Wciąż rośnie liczba obszarów, w których uczenie maszynowe znajduje zastosowania: od automatycznych tłumaczeń tekstów, poprzez samosterujące się pojazdy, gry planszowe oraz komputerowe, medycynę, aż do wszelkiego rodzaju zagadnień związanych z obróbką obrazów czy materiałów wideo. Ostatnie z wymienionych zastosowań (przetwarzanie materiałów graficznych oraz audio-wizualnych) szczególnie często wykorzystuje głębokie sieci splotowe, czyli specyficzny rodzaj sieci neuronowych, składających się z wielu warstw, wykorzystujących filtry splotowe do osiągania założonych celów (np. identyfikacji obiektów znajdujących się na zdjęciach czy wręcz tworzenia opisów tych obrazów w języku naturalnym).

Z powodu mnogości istniejących zastosowań dla sieci splotowych oraz prawdopodobnie wielu dziedzin wciąż czekających na wsparcie ze strony mechanizmów tego rodzaju, warto zgłębić tę szczególną gałąź uczenia maszynowego. Zadaniem referencyjnym, które jest często wykorzystywane do badań nad sieciami splotowymi jest identyfikacja przedmiotów przedstawionych na obrazkach. Istnieją również bazy opisanych etykietami zdjęć takie jak CIFAR-10 i CIFAR-100 [10] czy ImageNet [11], które mogą posłużyć jako zbiór danych treningowych i testowych.

W ramach swojej pracy magisterskiej stworzyłem głęboką splotową sieć neuronową realizującą wymienione zadanie referencyjne. Ze względu na niewielką moc obliczeniową urządzeń, z których mogłem korzystać, posłużyłem się bazą CIFAR-10. Pozwoliło to zbadać wpływ różnych parametrów sieci na jakość uczenia i klasyfikacji. Badano między innymi zachowanie sieci z różnymi algorytmami normalizacji, różnymi modelami neuronu oraz przedstawiono proces ulepszania sieci poprzez zmianę jej architektury. Wiedza ta pozwala konstruowanie większych sieci, w celu realizacji bardziej złożonych obliczeniowo zadań rozpoznawania obrazów.

Literatura.

[1] Andrej Karpathy. What a Deep Neural Network thinks about your #selfie. http://karpathy.github.io/2015/10/25/selfie/, 2015-10-25 [dostęp: 2016-12-15].

[2] David Garcia. Image Super Resolution. https://github.com/david-gpu/srez, 2016-10-18 [dostęp: 2016-12-15].

[3] Alec Radford, Luke Metz, Soumith Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. CoRR, abs/1511.06434, 2015.

[4] Google Deep Mind. AlphaGO. https://deepmind.com/research/alphago/ [dostęp: 2016-12-15].

[5] Google Deep Mind. In a Huge Breakthrough, Google's AI Beats a Top Player at the Game of Go. https://goo.gl/7Vq0kh [dostęp: 2016-12-15].

[6] Google Deep Mind. Google achieves AI 'breakthrough' by beating Go champion. http://www.bbc.com/news/technology-35420579, 2016-01-27 [dostęp: 2016-12-15].

[7] Jonathan Schaeffer, Neil Burch, Yngvi Björnsson, Akihiro Kishimoto, Martin Müller, Robert Lake, Paul Lu, Steve Sutphen. Checkers Is Solved, Science, vol. 317 no. 5844 1518-1522, 2007-07-06.

[8] Monroe Newborn. Kasparov versus Deep Blue : computer chess comes of age, Springer, 1997.

[9] Hugo Larochelle. Neural networks [7.1] : Deep learning – motivation. https://www.youtube.com/watch?v=vXMpKYRhpmI, 2013-11-15 [dostęp: 2016-12-15].

[10] Cifar-10 and Cifar-100 Datasets. https://www.cs.toronto.edu/~kriz/cifar.html [dostęp: 2016-12-15].

[11] Image Net. http://www.image-net.org/ [dostęp: 2016-12-15].