dr inż. Piotr Iwo Wójcik

Rozmowa z autorem pracy:
„Zastosowania metody rzutu przypadkowego w głębokich sieciach neuronowych"

Co przyciągnęło Pana do algorytmów sieci neuronowych? W jaki sposób Pan się nimi zainteresował?

Zacznijmy od tego, że sieci neuronowe są jednym z przykładów modeli uczenia maszynowego, a to właśnie uczenie maszynowe i sztuczna inteligencja interesują mnie najbardziej. Głównym zadaniem metod sztucznej inteligencji jest tworzenie programów będących w stanie realizować złożone, nieal-gorytmizowalne zadania, zazwyczaj wykonywane tylko przez ludzi, jak np. prowadzenie samochodu, granie w Go, Starcraft'a czy Quake'a, czy bardzo istotny problem identyfikowania kotów na zdjęciach. Czyli, w skrócie, sztuczna inteligencja to najciekawsze, co komputery potrafią obecnie robić.
Samym uczeniem maszynowym interesowałem się od dawna, natomiast moją uwagę na sieci neuronowe skierował promotor mojej pracy, prof. Witold Dzwinel, na początku studiów doktoranckich. Okazało się to bardzo dobrym wyborem, szczególnie w kontekście obecnej popularności i dynamicznego rozwoju sieci.

Jakie efekty dla uczenia maszynowego możemy uzyskać poprzez zmniejszenie wymiarowości danych wejściowych?

Redukcja wymiarowości pozwala przede wszystkim zmniejszyć koszt obliczeniowy oraz zużycie pamięci takich metod. W przypadku bardzo wymagających obliczeniowo modeli uczenia maszynowego, jak np. sieci neuronowe, redukcja wymiarowości wręcz umożliwia zastosowanie takich metod na danych zbyt dużych, by można było bezpośrednio trenować nimi model. Samo uczenie w niskowymiarowej przestrzeni cech wymaga mniej przykładów uczących, przez co może przebiegać szybciej. Dodatkowo, w zależności od metody redukcji wymiarowości, może mieć ona także wpływ regulujący, czyli ograniczać ryzyko zbytniego dopasowania się modelu do danych.

Dlaczego metoda rzutu przypadkowego sprawdza się w zakresie masywnych zbiorów danych o dużej ilości cech wejściowych?

Główny powód użyteczności metody rzutu przypadkowego do analizy olbrzymich zbiorów danych jest prosty: jest ona bardzo efektywna zarówno pod względem obliczeniowym jak i pamięciowym. Dodatkowo, może działać w trybie online, czyli rekord po rekordzie, bez konieczności wczytania całego zbioru do pamięci.

Co stanowiło dla Pana największe wyzwanie w czasie badań? Jakie przeszkody pojawiły się po drodze i jak sobie Pan z nimi poradził?

Największe wyzwanie wynikało z samej natury mojej pracy – zajmowałem się uczeniem modeli na bardzo dużych zbiorach danych. Wielkość tych zbiorów sprawiała, że konieczna była specjalna

infrastruktura umożliwiająca sprawne prowadzenie obliczeń. Na szczęście, Cyfronet dysponuje właśnie taką infrastrukturą. Warto zresztą dodać, że ograniczenie to powstrzymywało też innych badaczy przed eksploracją tego tematu, dzięki czemu wkraczałem ze swoją pracą na relatywnie nieprzebadany teren.
Natomiast osobiście dla mnie największą przeszkodą była konieczność rozdzielenia czasu pomiędzy pracę nad doktoratem a zobowiązaniami zawodowymi i po prostu innymi życiowymi aktywnościami. Z tą przeszkodą poradziłem sobie robiąc doktorat niemal dwa razy dłużej niż zakładane cztery lata :) . Nie żałuję tego jednak, gdyż fakt, że nie ograniczyłem się jedynie do pracy nad doktoratem, pozwolił mi zebrać wiele cennych doświadczeń, które wykorzystuję obecnie w pracy zawodowej.

Uczenie głębokich sieci neuronowych wymaga znacznego czasu obliczeniowego. Jakim wsparciem w tym zakresie są zasoby Cyfronetu?

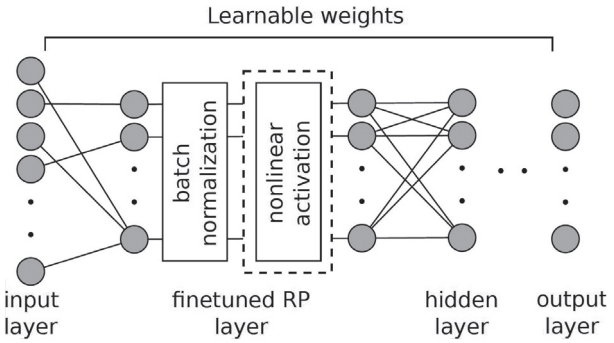
Nieocenionym. Ale żeby dać pewne wyobrażenie, policzmy: w trakcie swoich badań zużyłem około 200 tysięcy godzin obliczeniowych i jest to raczej wartość mocno niedoszacowana. W zdecydowanej większości były to zadania treningu sieci neuronowych intensywnie używające dostępnych na klastrze Prometheus kart graficznych Tesla K40 o mocy około 5 TFlops'ów każda. Gdyby te same obliczenia wykonywać na popularnej komercyjnej platformie Google Cloud Platform ML Engine i użyć podstawowej maszyny „standard_gpu" z ponad 50% szybszą kartą graficzną Tesla K80, obliczenia trwałyby prawdopodobnie ponad 125 tysięcy godzin. Obecnie cena jednej godziny treningu na maszynie „standard_gpu" kosztuje $0.93. Gdybym więc dzisiaj miał powtórzyć obliczenia na GCP, całkowity koszt wyniósłby ponad 400 tysięcy złotych. Czyli, efektywnie, bez wsparcia ze strony zasobów Cyfronetu nie mógłbym przeprowadzić swoich badań.

Jakie kolejne kroki mogą zostać podjęte na bazie osiągniętych przez Pana wyników?

Kontynuacja prac mogłaby iść w dwóch kierunkach. Po pierwsze, sama metoda mogłaby zostać udoskonalona, na przykład poprzez użycie niedawno zaproponowanych gęstych ustrukturyzowanych macierzy rzutu przypadkowego będących szczególnym przykładem macierzy Toeplitza. Drugim kierunkiem mogłoby być zastosowanie zaproponowanej metody w dziedzinach, gdzie wysoka wymiarowość i ilość danych uniemożliwia zastosowanie sieci neuronowych, jak np. do analizy danych genetycznych.

Na co powinny, Pana zdaniem, zwrócić uwagę osoby dopiero wkraczające na ścieżkę naukową?

Po pierwsze, chciałbym im zwrócić uwagę, że kariera naukowa, przynajmniej w zakresie uczenia maszynowego, nie jest nierozłącznie związana z uczelnią. Wiele firm ma obecnie świetne zespoły R&D, w ramach których można pracować i rozwijać się naukowo. Zespoły te publikują wyniki na najlepszych konferencjach i aktywnie uczestniczą w życiu naukowym, współpracując także z uczelniami.



Sieć neuronowa z douczaną warstwą rzutu przypadkowego

Niezależnie jednak, czy ktoś wybierze firmę czy uczelnię, sądzę, że kluczowy jest wybór właściwego dla siebie zespołu. Takiego, który rzeczywiście pozwala na rozwój, zapewnia wsparcie, ale też dobrą atmosferę. Myślę, że to porada, która sprawdza się w niemal każdej dziedzinie zawodowej.