dr hab. Piotr Wnuk-Lipiński
Zakład Inteligencji Obliczeniowej
Instytut Informatyki
Uniwersytet Wrocławski
ul. F. Joliot-Curie 15
50-383 Wrocław

Wrocław, dnia 14 stycznia 2019 r.

Recenzja rozprawy doktorskiej mgr. Stanisława Jastrzębskiego "Deep Networks Generalization and Optimization Trajectory"

Rozprawa doktorska dotyczy zrozumienia zjawiska uczenia się głębokich sieci neuronowych, zwłaszcza sposobów generalizowania i zapamiętywania wiedzy, oraz wpływu parametrów algorytmu uczącego na te zjawiska.

Formalnie, rozprawa doktorska składa się z 6 artykułów naukowych opublikowanych na międzynarodowych konferencjach naukowych poświęconych uczeniu maszynowemu i sieciom neuronowym, zebranych w całość, uzupełnionych o wstęp, omówienie i podsumowanie. W załączniku znalazły się dodatkowo wersje rozszerzone dwóch z opublikowanych artykułów.

Wszystkie artykuły naukowe wchodzące w skład rozprawy doktorskiej są wieloautorskie (co wydaje się standardem w tego typu badaniach naukowych), ale Autor rozprawy doktorskiej jest w każdym z nich wskazany jako autor główny (w 3 przypadkach jako jedyny autor główny).

1. Oryginalny wkład Autora zawarty w rozprawie doktorskiej

Najważniejsze osiągnięcia Autora zawarte w rozprawie doktorskiej dotyczą zrozumienia sposobu uczenia się sieci neuronowych algorytmem *Stochastic Gradient Descent* (SGD), zrozumienia wpływu parametrów algorytmu SGD, głównie kroku uczenia (ang. *learning rate*) i rozmiaru próbki (ang. *batch size*), na jego działania i otrzymywane rezultaty. W szczególności:

- Praca "*Width of Minima Reached by Stochastic Gradient Descent is Influenced by Learning Rate to Batch Size Ratio*" (P1) uzasadnia, że na krzywiznę funkcji kosztu znaczący wpływ ma krok uczenia i rozmiar próbki.

- Praca "*A Closer Look at Memorization in Deep Networks*" (P2) pokazuje, że głębokie sieci neuronowe mają tendencję do skupiania się na istotnych wzorcach i ich generalizowania, a nie zapamiętywania.

- Praca "*DNN's Sharpest Directions Along the SGD Trajectory*" (P3) pokazuje, że krzywizna funkcji kosztu ustala się w początkowej fazie uczenia, a odpowiednio dobierając parametry algorytmu uczącego można sterować kształtem tej krzywizny.

Uzyskane wyniki uważam za wartościowe, zarówno w kontekście akademickim, gdyż przyczyniają się do powiększenia wiedzy o zjawisku uczenia się sieci neuronowych, jak i praktycznym, gdyż umożliwiają lepszy dobór parametrów algorytmu uczenia sieci neuronowych, co przekłada się na zwiększenie ich efektywności.

Warto zwrócić uwagę, że praca P2 została opublikowana na bardzo dobrej międzynarodowej konferencji naukowej *International Conference on Machine Learning* poświęconej uczeniu maszynowemu, zaś praca P1 na dobrej międzynarodowej konferencji naukowej *International Conference on Artificial Neural Networks* poświęconej sieciom neuronowym.

2. Treść rozprawy doktorskiej

Rozdział 1 krótko przedstawia motywację prowadzonych badań naukowych dotyczących zrozumienia sposobu uczenia się sieci neuronowych, zwłaszcza sposobów generalizowania i zapamiętywania wiedzy, oraz wpływu parametrów na sposób generalizowania i zapamiętywania. Rozdział ten potwierdza, że Autor rozprawy jest świadomy celu swoich badań naukowych, potrafi umiejscowić je we współczesnym świecie nauki wiążąc z pracami innych naukowców, a także, że prace Autora wpisują się bardzo mocno w zainteresowania i w badania prowadzone w najlepszych ośrodkach naukowych na świecie w dziedzinie uczenia maszynowego (m.in. zespół Y. Bengio na University of Montreal, zespół S. Bengio w Google Brian, itd.).

Rozdział 2 wprowadza do tematyki uczenia maszynowego, wyjaśnia pojęcia generalizowania i zapamiętywania oraz opisuje algorytm SGD. Rozdział jest bardzo skromny (chociaż tłumaczy najważniejsze pojęcia używane w rozprawie doktorskiej). Zakłada nie tylko, że czytelnik posiada wiedzę z dziedziny uczenia maszynowego, ale także, że jest świadomy problemów związanych ze współczesnymi podejściami. Bynajmniej nie uważam tego za wadę - przyjęta forma, krótka i zwięzła, wydaje mi się stosowna do rozprawy doktorskiej komponowanej z opublikowanych artykułów naukowych.

Rozdział 3 dotyczy generalizowania, zapamiętywania i optymalizacji w uczeniu maszynowym. Streszcza osiągnięcia Autora opisane w pracach P1, P2 i P3, głównie w zakresie wpływu kroku uczenia i rozmiaru próbki na zachowanie algorytmu SGD. Rozdział 4 stanowi uzupełnienie poprzedniego rozdziału o streszczenia prac P4, P5 i P6, które dotyczą wpływu architektury i reprezentacji na generalizację i zapamiętywanie w uczeniu maszynowym. Rozdziały 3 i 4 są szczególnie cenne, bo Autor dyskutuje w nich całokształt wyników z perspektywy czasu, co pozwala na ocenę znaczenia całości przeprowadzonych badań naukowych.

Rozdział 5 podsumowuje uzyskane wyniki.

Rozdział 6 zawiera prace stanowiące sedno rozprawy doktorskiej.

W pracy "*Width of Minima Reached by Stochastic Gradient Descent is Influenced by Learning Rate to Batch Size Ratio*" (P1) autorzy analizują wpływ ilorazu kroku uczenia (ang. learning rate) i rozmiaru próbki (ang. batch size) na zachowanie algorytmu SGD, a w szczególności na jego dynamikę i zbieżność oraz na błąd generalizacji. Pokazano także związek między tym ilorazem a długością minimum wyznaczanego przez algorytm SGD. Eksperymenty obliczeniowe, potwierdzające rozważania teoretyczne, wykonano m.in. na sieciach ResNet i VGG11 dla CIFAR10 oraz ReLU MLP na Fashion-MNIST.

Badania wpisują się w coraz bardziej popularną tematykę studiów nad zachowaniem algorytmu SGD, badanych obecnie w wielu ośrodkach badawczych na świecie, m.in. przez zespoły Y. Bengio i S. Bengio, ale zapoczątkowanych jeszcze ubiegłym wieku, m.in. przez Schmidhubera w 1997. Uzyskane wyniki są wartościowe: oprócz przyczynku do większego zrozumienia sposobu uczenia się sieci neuronowych, mają także znaczenie praktyczne - mogą pomagać w doborze właściwych parametrów algorytmu SGD, co może ograniczyć eksperymentalne dobieranie tych parametrów i przyspieszyć uczenie sieci neuronowych.

Pewne niejasności dotyczące uniwersalności prezentowanego podejścia może powodować sposób walidacji podejścia, które autorzy pracy sprawdzali jedynie na problemach związanych z klasyfikacją i rozpoznawaniem obrazów (CIFAR10 i Fashion-MNIST). Wydaje się ciekawe sprawdzenie podejścia także na innych problemach, m.in. dotyczących rozpoznawania sygnałów, w szczególności rozpoznawania mowy, do czego głębokie sieci neuronowe są intensywnie używane w ostatnich latach.

Praca "*A Closer Look at Memorization in Deep Networks*" (P2) skupia się na mechanizmie generalizowania i zapamiętywania w głębokim uczeniu maszynowym. Autorzy pokazują, iż mimo, że głębokie sieci neuronowe są zdolne do zapamiętywania bardzo dużych ilości danych uczących (ze względu na bardzo dużą liczbę parametrów modelu), przejawiają tendencję do uczenia się prostych wzorców i generalizowania, a nie zapamiętywania. Praca rozszerza wcześniejsze badania naukowe prowadzone przez Zhanga i innych opublikowane w "*Understanding Deep Learning Requires Rethinking Generaization*" w 2017.

Autorzy wprowadzają pojęcie *krytycznej próbki danych* (ang. *critical sample*) i proponują algorytm *Langevin Adversarial Sample Search* (LASS) do identyfikacji takich próbek. Podczas uczenia głębokich sieci neuronowych, stosując odpowiednią regularyzację, zwiększają stopień uczenia się próbek krytycznych (co poprawia generalizowanie), a obniżają pozostałych (co osłabia zapamiętywanie). Eksperymenty obliczeniowe przeprowadzone głównie na zbiorach CIFAR10 i MNIST potwierdzają skuteczność proponowanego podejścia.

Innym ważnym osiągnięciem pracy wydaje się określenie wpływu architektury i sposobu uczenia głębokich sieci neuronowych na generalizowanie i zapamiętywanie oraz znaczenia regularyzacji.

Praca "*DNN's Sharpest Directions Along the SGD Trajectory*" (P3) uzupełnia prace P1 i P2. Autorzy skupiają się na kształcie minimum wyznaczanego przez algorytm SGD. Proponują podejście, w którym poprzez modyfikację algorytmu SGD, można sterować kształtem minimum otrzymując bardziej strome lub bardziej płaskie minima.

Otrzymane wyniki są interesujące i wartościowe, bo umożliwiają lepszą konfigurację algorytmów uczenia głębokich sieci neuronowych. Stanowią też istotne rozszerzenie wcześniejszych wyników dotyczących kształtu minimów, m.in. Schmidhubera i innych opublikowanych w "*Flat minima*" w 1997 oraz Dinha i innych opublikowanych w "*Sharp minima can generalize for deep nets*" w 2017.

Uzupełnieniem prac P1, P2 i P3 są prace "*Residual Connections Encourage Iterative Inference*" (P4), "*Commonsense Mining as Knowledge Base Completion? A Study on the Impact of Novelty*" (P5) i "*Learning to SMILE(S)*" (P6) uwzględniające znaczenie architektury i reprezentacji używanej w

głębokich sieciach neuronowych. W pracy P4 rozwinięto podejście Greffa i innych opublikowane w "*Highway and residual networks learn unrolled iterative estimation*" w 2016 dotyczące iteracyjnej estymacji w sieciach ResNet i pokazano podobieństwa do SGD. Praca P5 dotyczy uczenia maszynowego w kontekście *wiedzy zdroworozsądkowej* (ang. *commonsense knowledge*) i problemu zapamiętywania w takim uczeniu. Praca P6 dotyczy zastosowań uczenia maszynowego w komputerowo wspomaganym projektowaniu leków (ang. *Computer Assisted Drug Design*) do selekcji małego podzbioru molekuł z dużego zbioru kandydatów. Autorzy zaproponowali podejście SMILES oparte na sieciach neuronowych i reprezentacji tekstowej.

Najważniejsze ogólne osiągnięcia opublikowane w pracach P4, P5 i P6, to potwierdzenie, że architektura i reprezentacja w głębokich sieciach neuronowych ma duży wpływ na sposób uczenia się i generalizowanie.

Chociaż wysoko oceniam badania naukowe prowadzone przez Autora rozprawy doktorskiej, jak i jego publikacje, zastanawia mnie sposób walidacji proponowanych podejść. Po pierwsze, w większości eksperymenty obliczeniowe dotyczą zagadnień związanych z przetwarzaniem i rozpoznawaniem obrazów (m.in. zbiory danych CIFAR10, MNIST), a zupełnie pomijają zagadnienia związane choćby z przetwarzaniem dźwięku czy rozpoznawaniem mowy, które są obecnie równie istotnym zastosowaniem głębokich sieci neuronowych. Po drugie, zbiory danych CIFAR10 i MNIST są coraz częściej uznawane za dość proste dla głębokich sieci neuronowych, więc obecnie zastępowane są w eksperymentach obliczeniowych przez wiele innych (m.in. CIFAR100, SVHN, itp).

3. Redakcja rozprawy doktorskiej

Rozprawa doktorska składa się z 6 rozdziałów i dodatku. Początkowe rozdziały stanowią wprowadzenie do opublikowanych artykułów naukowych, streszczają najważniejsze osiągnięcia i przedstawiają całość w szerszym świetle z perspektywy czasu (co zapewne nie było możliwe podczas pisania pojedynczych artykułów). Rozdział 6 zawiera 6 artykułów naukowych stanowiących główne osiągnięcie naukowe prezentowane w rozprawie doktorskiej, zaś dodatek zawiera rozszerzone wersje artykułów P1 i P3.

Forma rozprawy doktorskiej wydaje się odpowiednia do treści i stosowna dla rozprawy doktorskiej. Warto jednak zauważyć, że praca jest dość hermetyczna - zdecydowanie skierowana do czytelnika posiadającego dość dużą wiedzę z dziedziny uczenia maszynowego i świadomego współczesnych problemów tej dziedziny. Czytelnik dysponujący jedynie podstawową wiedzą matematyczną i informatyczną będzie zmuszony sięgnąć po dodatkową literaturę z dziedziny uczenia maszynowego. Nie uważam tego za wadę rozprawy doktorskiej.

Praca jest zrozumiała i napisana w sposób staranny. Szczególnie spodobał mi się sposób połączenia opublikowanych artykułów naukowych z początkowymi rozdziałami rozprawy doktorskiej (wstępem, omówieniem i podsumowaniem), co uczyniło całość rozprawy spójną i przyjemną do czytania.

4. Konkluzja

Moje sugestie i uwagi krytyczne wymienione w treści niniejszej recenzji, dotyczące wskazanych fragmentów pracy, nie zmieniają mojej pozytywnej opinii o rozprawie doktorskiej.

Uważam, że rozprawa doktorska mgr. Stanisława Jastrzębskiego stanowi oryginalne rozwiązanie problemu naukowego, wykazuje ogólną wiedzę teoretyczną kandydata w danej dyscyplinie naukowej oraz umiejętność samodzielnego prowadzenia pracy naukowej i spełnia warunki określone w *Ustawie z dnia 14 marca 2003 o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuk* (Dz.U. 2003, nr 65, poz. 595, z późniejszymi zmianami). Wnoszę więc o dopuszczenie rozprawy doktorskiej mgr. Stanisława Jastrzębskiego do publicznej obrony.

Biorąc pod uwagę oryginalny i nowatorski charakter badań naukowych przedstawionych w rozprawie doktorskiej mgr. Stanisława Jastrzębskiego oraz znaczenie podjętego tematu, wnioskuję o wyróżnienie pracy przez Radę Wydziału Matematyki i Informatyki UJ.

Piotr Lipiński