dr hab. inż. Krzysztof Dembczyński
Instytut Informatyki (Institute of Computing Science)
Politechnika Poznańska (Poznań University of Technology)
ul. Piotrowo 2, 60-965 Poznań
tel: (+48) 61 665 2936
kdembczynski@cs.put.poznan.pl

4 lutego 2021 roku

Recenzja rozprawy doktorskiej
Tytuł: Robustness of Neural Networks
Autor: Konrad Żołna

1 Krótkie podsumowanie i streszczenie rozprawy

Rozprawa dotyczy problemu uczenia sztucznych sieci neuronowych w sposób zapewniający odporność na zwodnicze (złośliwe) przykłady uczące. Autor proponuje szereg rozwiązań luźno ze sobą powiązanych dla szerokiego spektrum problemów uczenia maszynowego. Rozprawa została złożona jako zbiór sześciu publikacji naukowych poprzedzony 15-stronicowym streszczeniem. Wszystkie przedstawione publikacje oraz streszczenie napisane są w języku angielskim. Poniżej przedstawiam dane bibliograficzne publikacji:

- Zajac, M., Żołna, K., Rostamzadeh, N., i Pinheiro, P. O. (2019). Adversarial framing for image and video classification. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 10077–10078. AAAI Press
- Żołna, K., Geras, K. J., i Cho, K. (2020). Classifier-agnostic saliency map extraction. *Comput. Vis. Image Underst.*, 196:102969
- Stachura, D., Galias, C., i Żołna, K. (2020). Leakage-robust classifier via mask-enhanced training (student abstract). In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 13923–13924. AAAI Press
- Ke, N. R., Żołna, K., Sordoni, A., Lin, Z., Trischler, A., Bengio, Y., Pineau, J., Charlin, L., i Pal, C. J. (2018). Focused hierarchical RNNs for conditional sequence processing. In Dy, J. G. i Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2559–2568. PMLR
- Żołna, K., Arpit, D., Suhubdy, D., i Bengio, Y. (2018). Fraternal dropout. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net
- Żołna, K. (2017). Improving the performance of neural networks in regression tasks using drawering. In *2017 International Joint Conference on Neural Networks, IJCNN 2017, Anchorage, AK, USA, May 14-19, 2017*, pages 2533–2538. IEEE

Niestety autor nie przedstawia bezpośrednio tezy rozprawy, co jest mocno zaskakujące. Z treści rozprawy można wywnioskować, że autor stara się traktować odporność jako bardzo ogólny problem generalizacji podczas uczenia z niedoskonałych danych zawierających przykłady złośliwe,

zaszumione lub błędne. Proponowane rozwiązania dotyczą jednak specyficznych problemów i rodzajów niedoskonałości danych. Nie zmienia to faktu, że zaprezentowane wyniki są oryginalne oraz dotyczą bardzo aktualnych problemów.

Streszczenie przedstawia ogólny zarys rozważanego zagadnienia oraz zawiera opis każdej publikacji, składający się z sformułowania problemu, stanu wiedzy, zaproponowanego rozwiązania oraz najważniejszych wyników. Należy przyznać, że streszczenie jest napisane ciekawie, zrozumiale opisuje otrzymane wyniki.

Pierwsza z zawartych prac dotyczy generowania przykładów złośliwych w problemie klasyfikacji obrazów. Autorzy tej pracy pokazują, że możliwe jest "zmylenie" klasyfikatora poprzez dodanie niewielkiej ramki do obrazu. Ramki o szerokości 4 pikseli są w stanie całkowicie zmylić klasyfikator. W drugiej pracy została zaproponowana metoda ekstrakcji najbardziej informatywnej części obrazu, która może wspomóc podejmowanie decyzji przez człowieka, np. lekarza przy analizie obrazów medycznych. Ponadto praca ta pokazuje, że sieci neuronowe skupiają się na specyficznych detalach obrazów, przez co mogą one być łatwym celem ataków. Powyższa obserwacja została dogłębniej przeanalizowane w kolejnym artykule. Na jej podstawie został zaproponowany nowy algorytm uczący. Czwarta praca proponuje nową architekturę rekurencyjnych sieci neuronowych. W szczególności zaproponowany został model, który dzieli sekwencję wejściową na rozłączne podciągi i podsumowuje każdy z nich. Oprócz tego dodatkowy moduł jest trenowany równolegle na reprezentacjach podciągów w celu rozwiązania danego problemu. Piąta praca wprowadza nową technikę regularyzacji, będącą specyficzną odmianą podejścia "dropout". Stosuje ona dwie maski zerujące parametry sieci podczas każdej iteracji trenowania. Różnica w predykcji jest następnie brana pod uwagę przy aktualizacji parametrów sieci. Ostatnia praca dotyczy problemu regresji. Autor pokazuje, że trenowanie sieci z dodatkową dyskretną warstwą wyjściową może skutecznie poprawić parametry wewnętrznych warstw sieci, tak że oryginalna warstwa wyjściowa otrzymuje lepsze wyniki dla problemu regresji. Ta dodatkowa warstwa wyjściowa jest uczona na uporządkowanych etykietach, otrzymanych poprzez dyskretyzację oryginalnej zmiennej ciągłej.

Wszystkie artykułu napisane są bardzo dobrze. Ich charakter jest głównie eksperymentalny, jednak zaproponowane algorytmy są opisane formalnie. Widać po nich, że autor rozprawy ma bardzo dobrą intuicję badawczą, posiada dobre zrozumienie matematyczne omawianych problemów, oraz szeroką wiedzę dotyczącą współczesnych osiągnięć w obszarze sztucznych sieci neuronowych.

2 Uwagi i komentarze o charakterze ogólnym

Rozprawa dotyczy bardzo aktualnych i ciekawych zagadnień dotyczących uczenia odpornych sztucznych sieci neuronowych. Spektrum analizowanych problemów oraz zaproponowanych metod jest bardzo szerokie. Wszystkie publikacje będące częścią rozprawy zostały opublikowane w prestiżowych i uznanych wydawnictwach, takich jak czasopismo *Computer Vision and Image Understanding* oraz na konferencjach *Interntational Conference on Machine Learning* (ICML), *AAAI Conference on Artificial Inteligence* (AAAI), *International Conference on Learning Representations* (ICLR), czy *International Joint Conference on Neural Networks* (IJCNN). Przede wszystkim na uwagę zasługują artykuły, które zostały przyjęte na konferencje ICML, AAAI oraz ICLR, należące do najważniejszych w dziedzinie uczenia maszynowego. Mojego uznania nie zmienia fakt, że dwie prace z konferencji AAAI mają status rozszerzonego streszczenia prac studenckich. Warto również dodać, że zawarte w rozprawie artykuły nie są jedynymi tak wysoko opublikowanymi pracami autora. Każdemu doktorantowi życzyłbym takich sukcesów. Należy jednak zauważyć, że tylko jedna praca jest napisane samodzielnie przez autora. Pozostałe prace są wynikiem współpracy z wieloma osobami, czasem z pokaźną ich liczbą. Warto jednak podkreślić, że autor współpracuje z bardzo znanymi grupami badawczymi, co również zasługuje na uznanie, ponieważ badania naukowe to także umiejętność współpracy.

Każda recenzja musi jednak ocenić pracę w sposób obiektywny, zauważając jej pozytywne i negatywne strony. Tak jak zostało wspomniane wcześniej, rozprawa nie przedstawia bezpośrednio tezy, której obrony podejmuje się autor. Z osobistego doświadczenie wiem, że sformułowanie tezy jest trudne i czasem wydaje się sztuczne. Jednak z perspektywy recenzenta postawienie tezy jest kluczowe, ponieważ pozwala łatwiej ocenić oraz podsumować osiągnięcia autora. Stąd rozprawa wydaje się luźnym połączeniem sześciu prac, bez całościowego omówienia problemu uczenia odpornych sztucznych sieci neuronowych. Część wyników zawartych w przedstawionych pracach wymaga

pełniejszego opisu i lepszego powiązania. Rozprawa doktorska jest właśnie tym miejscem, w którym można odpowiednio rozszerzyć wcześniejsze prace i powiązać je ze sobą. W pełni wartościowa rozprawa powinna stanowić spójną całość, omawiającą wyczerpująco dany problem badawczy, niekoniecznie w zakresie tak szerokim jak przedstawione prace autora. Jeżeli przynajmniej jedna z załączonych prac byłaby próbą podsumowania wcześniejszych badań, wtedy rozprawa w formie zbioru artykułów byłoby bardziej uzasadniona. Niestety przedstawiona rozprawa sprawia poczucie niedosytu, jednak nie ujmującego osiągnięć autora.

3 Uwagi szczegółowe

Przedstawione prace są napisane bardzo dobrze i przeszły już surowy proces recenzyjny. Stąd lista moich szczegółowych uwag jest krótka.

- Zastanowiające jest to, że autor dwukrotnie błędnie wskazuje miejsce publikacji jednej z prac. Zamiast *International Conference on Learning Representations* pojawia się *International Conference on Learning Algorithms*. Mam nadzieję, że autor jest świadomy na jakiej konferencji została zaprezentowana jego praca :).
- Czy "złośliwa ramka" jest taka sama dla wszystkich obrazów? Czy atak z wieloma wariantami "ramki" nie byłby bardziej skuteczny?
- W pracy *Classifier-agnostic saliency map extraction* używana jest metoda próbkowania funkcji. Przypomina ona podejście zwane próbkowaniem zbiornikowym (ang. *reservoir sampling*), jednak opis zastosowanego algorytmu nie jest dostatecznie czytelny. Jaka jest relacja pomiędzy zastosowaną metodą oraz próbkowaniem zbiornikowym?
- W tej samej pracy zdefiniowany jest problem optymalizacji w celu otrzymania mapy istotności, który jest bardzo kosztowny. Następnie autorzy prezentują algorytm, który w sposób przybliżony rozwiązuje oryginalny problem. Czy jakość tego przybliżenia została zweryfikowane teoretycznie lub empirycznie?
- Praca *Focused Hierarchical RNNs for Conditional Sequence Processing* została napisana przez 9-ciu autorów. Jaki dokładnie wkład w tę pracę miał autor rozprawy?
- W ostatniej pracy problem regresji jest rozwiązywany przy pomocy dodatkowej warstwy wyjściowej uczonej na podstawie uporządkowanych etykiet wieloklasowych. Taki problem nazywany jest klasyfikacją lub regresją porządkową. Niestety w pracy nie ma bezpośredniego nawiązania do tego problemu. Czy zastosowanie algorytmów klasyfikacji porządkowej wpłynęłoby korzystnie na rozwiązanie oryginalnego problemu regresji?

4 Konkluzja końcowa

Przedstawiona rozprawa prezentuje imponujące wyniki dotyczące szerokiego spektrum problemów związanych z uczeniem odpornych sztucznych sieci neuronowych. Moje krytyczne komentarze dotyczące uspójnienia przedstawionych wyników i zaprezentowania ich w formie jednolitego tekstu nie wpływają na moją wysoką ocenę osiągnięć naukowych autora rozprawy.

Podsumowując powyższą recenzję wnoszę o dopuszczenie Pana mgr. Konrada Żołny do dalszych etapów przewodu doktorskiego. Jestem także skłonny poprzeć wniosek o wyróżnienie rozprawy, jeżeli zostałby on zgłoszony przez drugiego recenzenta.

dr hab. inż. Krzysztof Dembczyński