UNIWERSYTET
WARSZAWSKI
Wydział Matematyki, Informatyki i Mechaniki
Instytut Informatyki

Warszawa, 13.11.2020

Recenzja rozprawy doktorskiej magistra Konrada Żołny pod tytułem *Robustness of Neural Networks*

Artykuły na których bazuje rozprawa

(1) M. Zając, K. Żołna, N. Rostamzadeh, P. O. Pinheiro. *Adversarial Framing for Image and Video Classification (Student Abstract)*. AAAI Conference on Artificial Intelligence. 2019.
(2) K. Żołna, K. J. Geras, K. Cho. *Classifier-Agnostic Saliency Map Extraction.* Computer Vision and Image Understanding. 2020.
(3) D. Stachura, C. Galias, K. Żołna. *Leakage-Robust Classifier via Mask-Enhanced Training (Student Abstract)*. AAAI Conference on Artificial Intelligence. 2020.
(4) N. R. Ke, K. Żołna, A. Sordoni, Z. Lin, A. Trischler, Y. Bengio, J. Pineau, L. Charlin, C. Pal. *Focused Hierarchical RNNs for Conditional Sequence Processing.* International Conference on Machine Learning. 2018.
(5) K. Żołna, D. Arpit, D. Suhubdy, Y. Bengio. *Fraternal Dropout.* International Conference on Learning Algorithms. 2018.
(6) K. Żołna. *Improving the Performance of Neural Networks in Regression Tasks Using Drawering*. International Joint Conference on Neural Networks. 2017.

Tematyka rozprawy

W ostatnich latach miało miejsce wiele przełomowych wyników wykorzystujących sieci neuronowe w dziedzinach takich jak rozpoznawanie obrazów, automatyczne tłumaczenie czy też gra w Go. Rozprawa wpisuje się w tę dynamicznie rozwijającą się dziedzinę, gdyż jej tematem jest niezawodność modeli opartych o sieci neuronowe w kontekstach takich jak zwodnicze przykłady (ang. *adversarial examples*), interpretowalność wyników, metody regularyzacji i nowe architektury dla sieci rekurencyjnych.

Wyniki rozprawy

W pracy (1) główny pomysł polega na wstawieniu do danego obrazu (lub strumienia wideo) ramki o grubości kilku pikseli, która to ramka ma powodować, że obraz (lub wideo) będzie zwodniczym przykładem - zadana sieć neuronowa będzie się mylić pomimo iż środek obrazu pozostał w niezmienionej formie. Co ciekawe, autorzy wykazali, że dla ustalonego modelu (zestawu wag) możliwe jest opracowanie jednej ramki, która będzie wykorzystywana w takiej samej formie dla wszystkich przetwarzanych obrazów.

Celem pracy (2) jest opracowanie modelu pozwalającego na wykrywanie istotnych części danego obrazu (ang. *saliency maps*). Co istotne, autorów interesuje przypadek w którym rzeczony model jest niezależny od wytrenowanej sieci neuronowej a także od klasy którą przedstawia zadany obraz. W istocie opracowany model pozwala na wykrywanie części obrazu istotnych z punktu widzenia *dowolnego* klasyfikatora, dzięki czemu generowane wzorce są bardzo często spójne.

Dodatkową zaletą przedstawionego podejścia jest możliwość lokalizacji szukanego obiektu na zdjęciu za pomocą znalezienia prostokąta okalającego największy spójny istotny fragment zadanego zdjęcia. Taki lokalizator powstaje z modelu rozwiązującego problem klasyfikacji, nie mający dostępu do prawidłowych lokalizacji szukanych obiektów nawet w zbiorze treningowym. Dla tak określonego problemu trenowania ze słabym nadzorem (ang. *weak supervised learning*) autorzy uzyskali najlepsze znane wyniki na zbiorze ILSVRC'14.

Aby uzyskać przedstawione wyniki autorzy użyli pomysłowego schematu trenowania modelu wykrywającego istotne części zdjęcia naprzemiennie z trenowaniem ciągu klasyfikatorów. Dzięki temu w każdym kroku uczenia generatora używany jest nowo losowany klasyfikator, podobnie jak w przypadku Generative Adversarial Networks. Co więcej, autorzy zaproponowali rozbudowaną funkcję celu, która odpowiednio balansuje czułość i precyzję. Podsumowując, praca zawiera nie tylko bardzo dobre wyniki ilościowe i jakościowe, jak również wymagała wielu posłów i technicznej sprawności w warstwie matematycznej.

Pracę (3) można traktować jako pewnego rozszerzenie pracy (2), gdyż w pracy (3) celem jest opracowanie schematu uczenia sieci neuronowych tak aby w sytuacji wycieku informacji (ang. *data leakage*) wytrenowany model nie koncentrował się na pikselach zawierających wyciek, a wykorzystywał możliwie dużo dostępnych cech występujących w wejściowym obrazie. Przedstawiony schemat uczenia polega na użyciu generatora istotnych części obrazu z pracy (2) - tym razem trenowanego pod kątem konkretnego klasyfikatora, zaczernieniu fragmentów obrazu odpowiadającym najistotniejszym z punktu widzenia klasyfikatora pikseli i dotrenowaniu modelu tak aby klasyfikował niepełny obraz. W ten sposób podczas trenowania sieć uczy się rozpoznawać obrazy z zasłoniętą częścią zawierającą wyciek danych.

Głównym wynikiem w pracy (4) jest nowa architektura rekurencyjnej sieci neuronowej, która przetwarzając kolejne elementy ciągu podejmuje nieodwracalną decyzję czy dany element powinien być brany pod uwagę w dalszych obliczeniach. Żeby móc taką decyzję podjąć model ma do dyspozycji nie tylko sekwencyjnie podany ciąg, ale także pytanie na które trenowany model ma odpowiedzieć (reprezentowane jako wynik funkcji zanurzającej) - na przykład model jako wejście ma konkretne pytanie oraz artykuł w Wikipedii z którego ma wydobyć odpowiedź.

Warto zwrócić uwagę, że decyzja czy model ma wziąć pod uwagę dany element ciągu jest binarna, a w związku z tym nie można zastosować standardowego podejścia przez propagację wsteczną stosowaną w uczeniu z nadzorem. Z tego powodu autorzy wykorzystują metody uczenia ze wzmocnieniem (ang. *reinforcement learning*). Ponadto, aby uniknąć sytuacji, w której model będzie zawsze wykorzystywał wszystkie elementy ciągu autorzy stosują regularyzację w postaci kary za każdy wykorzystany element ciągu.

Prezentowana metoda z pracy (4) jest eksperymentalnie oceniana na zbiorach MS MARCO oraz SearchQA, gdzie osiągane są wyniki porównywalne lub lepsze od aktualnego stanu wiedzy. Ponadto analizowane są prostsze zadania (takie jak wybór najczęściej występującego elementu spośród prefiksu długości k), gdzie autorzy pokazują, że model ma własności zgodne z oczekiwaniami i prezentowaną narracją.

W pracy (5) autorzy przedstawiają nowy sposób regularyzacji, który jest pewnym wariantem metody dropout. W klasycznej metodzie dropout podczas treningu neurony są wyłączane w sposób losowy, natomiast podczas ewaluacji modelu na zbiorze testowym używane są wszystkie neurony lecz ich wartości są przeskalowane o czynnik gwarantujący utrzymanie średniej wartości oczekiwanej obliczanej przez każdy z neuronów.

Głównym wynikiem pracy jest metoda w której sieć trenowana jest na parach egzemplarzy danych dla dwóch różnych masek binarnych dla funkcji dropout, a dodatkowo stosowana jest nowy składnik w optymalizowanej funkcji straty, który to składnik odpowiada za minimalizację rozbieżności wektorów cech generowanych przez model dla wspomnianych dwóch różnych masek binarnych.

Prezentowana metoda jest eksperymentalnie ewaluowana w dwóch kontekstach:

- jako forma regularyzacji sieci rekurencyjnych, gdzie efekty są testowane na zbiorach Penn Tree-bank oraz WikiText-2, osiągając najlepsze znane wyniki.
- jako funkcja celu w przypadku nieetykietowanych danych dla problemu klasyfikacji w podejściu półnadzorowanym (ang. *semi supervised*).

Ostatnia z cyklu prac (6) dotyczy rozszerzenia sieci neuronowych rozwiązujących problem regresji o dodatkową głowicę rozwiązującą problem klasyfikacji. Zadany przedział w problemie regresji jest podzielony na podprzedziały, które pokrywają cały zakres możliwych odpowiedzi. Autorzy rozważają różne warianty podziału, gdzie przedziały są równe bądź nierówne, rozłączne lub zachodzące na siebie. Dodatkowo problem klasyfikacji może przybierać formę odpowiedzi na pytanie czy szukana liczba należy do przedziału [a,b] lub też czy szukana liczba jest nie mniejsza niż a. Warto wspomnieć, że aby uniknąć problemu różnych skal gradientów płynących z dwóch składników funkcji celu, autorzy proponują metodę ich balansowania.

Prezentowana metoda jest ewaluowana między innymi na zbiorze danych sieci sklepów Rossman, zbiór ten dostępny jest na platformie *kaggle.com*. Weryfikacja eksperymentalna wykazała poprawę wyników na omawianym zbiorze danych, co autorzy przypisują lepszym własnościom gradientów w przypadku problemu klasyfikacji.

Ocena rozprawy

W moim przekonaniu przedstawiona rozprawa jest na bardzo wysokim poziomie merytorycznym w skali światowej. Ponadto jest ona spójna tematycznie. Prezentowany cykl prac zawiera szereg nowatorskich pomysłów i bardzo dobrze wpisuje się badania bardzo prężnej i aktualnej dziedziny.

Spośród prezentowanego cyklu, prace (2), (4), (5) uznaję za najbardziej wartościowe, co również odzwierciedla renoma konferencji na których były prezentowane (ICML, AAAI, ICLR). Ponadto doceniam elegancję i prostotę pomysłów omawianych w pracach (1) oraz (3). Z technicznego punktu widzenia prace (2) oraz (4) wymagały bardzo dobrego rzemiosła zarówno od strony inżynierskiej (nietrywialna implementacja) jak również matematycznej. Pracę (6) uznaję za mniej nowatorską, gdyż pomysły w niej omawiane były znane w środowisku konkursów uczenia maszynowego, jednakże z pewnością praca stanowi wartość dodaną dla środowiska, a na uwagę zasługuje fakt iż jest to praca napisana samodzielnie przez autora i to na wczesnym etapie studiów doktoranckich.

Podsumowanie

W mojej ocenie przedstawiona rozprawa doktorska i dorobek naukowy magistra Konrada Żołny bez cienia wątpliwości spełnia wymagania ustawowe i zwyczajowe stawiane rozprawom doktorskim, a co za tym idzie rozprawa ta może stanowić podstawę do nadania tytułu doktora. Co więcej, rekomenduję wyróżnienie rozprawy.

Marek Cygan

Marek Cygan