dr hab. inż Jan Chorowski
Instytut Informatyki Uniwersytetu Wrocławskiego
ul. F. Joliot-Curie 15
50-383 Wrocław

Wrocław, 15.10.2020

Recenzja rozprawy doktorskiej mgr Konrada Żołny pt. „Robustness of Neural Networks"

1 Tematyka rozprawy

Praca składa się z 6 powiązanych tematycznie artykułów podejmujących temat niezawodności sieci neuronowych, rozumianej jako odporności sieci na zaburzenia, w tym zaburzenia powodujące że dane uczące i testowe pochodzą z różnych rozkładów. W przytoczonych pracach przedstawiono metody:

- Generowania adwersaryjnych obrazów i filmów przez dodanie ramki powodującej kontrolowane zaburzenie wyników sieci.
- Metodę określania dla klasyfikowanych obrazów ich istotnych pikseli.
- Wykorzystanie powyższej metody do regularyzacji modeli przez pozbawienie ich możliwości wykorzystywania tylko małych fragmentów wejścia do podjęcia decyzji.
- Metodę selektywnego skracania danych sekwencyjnych, przez zastosowanie mechanizmu binarnych bramek.
- Regularyzację sieci neuronowych przez wymuszenie spójnego działania sieci przy stosowaniu różnych masek dropout.
- Metodę rozszerzania sieci rozwiązujących zadanie regresji o dodatkowe zadanie klasyfikacji, wspierające regularyzację sieci.

Zestawienie artykułów pokazuje na kompleksowe podejście Doktoranta do tematu: przedstawiono metody diagnostyki działania sieci, oraz metody regularyzacji umożliwiające przekazanie sieci naszej wiedzy dziedzinowej o problemie (np. przez założenie że tylko część pikseli klasyfikowanych obrazów, lub tylko część słów w klasyfikowanym paragrafie jest istotna).

2 Ocena rozprawy

Rozprawa składa się z 6 powiązanych tematycznie artykułów, do których zamieszczono wspólny wstęp oraz streszczenie wyników widziane przez pryzmat tematu rozprawy. Następnie przedrukowano treści prac. Wszystkie prace opublikowano w bardzo dobrym czasopiśmie i na bardzo dobrych konferencjach:

- 1 artykuł w czasopiśmie Computer Vision and Image Understanding,
- 1 Artykuł na Int. Conf. on Machine Learnig,
- 1 artykuł na Int. Conf. on Learning Representations,
- 2 extended abstracts na konf. AAAI,
- i jeden artykuł na Int. Joint Conf. on Neural Networks.

Dodatkowo, dla części prac upubliczniono kody źródłowe umożliwiające replikację wyników.

Na uwagę zasługuje mnogość współpracy naukowych nawiązanych przez Doktoranta podczas studiów na Uniwersytecie Jagielońskim. Przytoczone prace powstawały z udziałem badaczy z Montrealskiego instytutu MILA, Uniwersytetu Nowojorskiego, oraz przedsiębiorstw.

2.1 Ocena treści

Rozdział 2. pracy zawiera przedstawienie prac zebranych w rozprawie. W sposób zwięzły, ale precyzyjny Autor przedstawił dla każdej z prac jej główne tezy, niezbędne informacje wstępne, uzyskane wyniki oraz ich znaczenie dla przewodniego tematu rozprawy.

Tak przygotowany wstęp pozwala ocenić rozmiar dorobku Doktoranta, zrozumieć sposób w jaki formułuje tezy naukowe oraz dostrzec logiczny ciąg podejmowanych działań.

Sekcja 2.1 oraz praca „Adversarial Framing".
W pracy zrealizowano atak na sieci neuronowe polegający na dodaniu do klasyfikowanego obrazu ramki o szerokości kilku pikseli. Wygenerowana ramka jest uniwersalna, taka sama dla każdego klasyfikowanego obrazu. To powoduje, że system może być stosowany np. do mylenia klasyfikatorów filtrujących treści w serwisach przechowujących filmy. Ograniczeniem proponowanego modelu jest założenie dostępu do oszukiwanego klasyfikatora. Autor wskazuje jednak na prace wykazujące że pewne obrazy adwersaryjne skutecznie manipulują działaniem różnych klasyfikatorów uczonych na podobnych danych.

Sekcja 2.2 oraz praca „Classifier-Agnostic Saliency Map Extraction".
W pracy podjęto temat analizy odpowiedzi sieci klasyfikujących obrazy, konkretnie możliwości wskazania fragmentu obrazu (maski pikseli) mających największy wpływ na odpowiedź klasyfikatora. Głównym założeniem pracy jest stworzenie metody uniwersalnej, niezależnej od instancji klasyfikatora którego odpowiedź jest tłumaczona. W tym sensie metoda znajduje istotne cechy danych.

Procedura CASME wykorzystuje kilka istotnych pomysłów, niektórych bardzo nowych:

- Ekstraktor maski m jest siecią neuronową, kategoryzującą każdy piksel jako istotny bądź nieistotny.
- Zbiór wszystkich możliwych parametryzacji danej sieci klasyfikującej jest niemożliwy do uzyskania. Zamiast tego, wykorzystano próbkowanie Langevina: ekstraktor uczono na różnych wagach klasyfikatorach z różnych iteracji nieskończonego treningu SGD.
- Trening ekstraktora przypomina uczenie GAN – maska nieistotnych pikseli ma maksymalnie zmylić klasyfikator.
- Uczenie ekstraktora ma charakter adwersaryjny i może generować maski mało semantyczne maski, wykorzystujące słabość klasyfikatora. Aby tego uniknąć konieczne jest ciągłe dotrenowywanie klasyfikatora na zamaskowanych obrazach.
- Klasyfikator i ekstraktor maski współdzielą parametry – ekstraktor architektura przypomina sieć typu U-Net, współdzielącą enkoder z klasyfikatorem.

Wyniki procedury są bardzo dobre. Poprawnie lokalizowany jest obiekt pierwszoplanowy. Maski poprawnie pogarszają wyniki klasyfikacji zarówno sieci współdzielących architekturę enkoder, jak i innych sieci dostępnych w bibliotece pytorch. Ponadto pozytywnie zweryfikowano możliwość maskowania obiektów innych klas niż te na których model był uczony.

Artykuł stanowi ważny wkład w prezentowaną rozprawę doktorską z dwóch powodów. Po pierwsze, przedstawia on bardzo dobrze wykonaną pracę zawierającą bardzo użyteczne i praktycznie stosowalne narzędzie. Po drugie, ilość stosowanych przez autorów technik oraz dyskusja doboru hiperparametrów CASME pokazują, że Doktorant dobrze opanował i doskonale rozumie techniki uczenia głębokiego.

Sekcja 2.3 oraz praca „Leakage-Robust Classifier."

W przedstawionym rozszerzonym abstrakcie wykorzystano metodę ekstrakcji masek pikseli CASME opisaną w poprzedniej pracy do zmuszenia klasyfikatora do wykorzystywania maksymalnie wielu informacji. Zaraportowany wyniki jednego syntetycznego eksperymentu: każdy obraz w zbiorze uczącym przetworzono, kodując poprawną klasę na kilku pikselach. W ten sposób uzyskano zbiór danych na którym nawet prosty klasyfikator mógł osiągnąć wysokie dokładności przez analizę jedynie specjalnie przygotowanych pikseli. Następnie wykazano, że klasyfikator uczony z maskowaniem małej liczby najbardziej informatywnych pikseli skutecznie ignoruje skrót i uczy się wnioskować o obrazach z pominięciem pikseli sztucznie kodujących klasę.

Uzyskany wynik na sztucznie zaburzonych danych pozostawia niedosyt i zachęca do dalszych badań, tym razem na danych rzeczywistych. Intrygujące byłoby np. wskazanie czy maskowanie tła poprawi klasyfikację tzw. naturalnych obrazów adwersaryjnych[1].

[1] https://arxiv.org/abs/1907.07174

Sekcja 2.4 oraz praca „Focused Hierarchical RNNs".

W pracy przedstawiono architekturą dwu-warstwowej sieci LSTM w której neurony w warstwie wyższej wykonują mniejszą ilość kroków niż neurony w warstwie niższej. Przejścia w warstwie wyższej są kontrolowane przez specjalne bramki binarne, uwarunkowane na zadaniu rozwiązywanym przez sieć.

Proponowana dwuwarstwowa sieć LSTM jest testowana na szeregu zadań, w tym na sztucznych zadaniach mających zweryfikować poprawność zaimplementowanego mechanizmu oraz na zadaniu odpowiedzi na pytania na podstawie załączonego paragrafu tekstu. Wykazano, że proponowana zmiana poprawia zdolność sieci do generalizowania na zdania dłuższe niż te widziane podczas treningu.

Proponowana architektura zakłada podejmowanie przez sieć dyskretnych decyzji – warstwa wyższa wybiera czy ma wykonać przejście, czy nie. Wymagało to zastosowanie metod estymacji gradientu uczącego zaczerpniętych z technik uczenia ze wzmocnieniem.

Sekcja 2.5 oraz praca „Fraternal Dropout".

W pracy zaproponowano regularyzację rekurencyjnych sieci neuronowych polegającą na wymuszaniu podobieństwa aktywacji sieci przy zastosowaniu dwóch różnych masek dropout. Eksperymentalnie wykazano, że stosowanie tej formy dropout poprawia działanie sieci rekurencyjnych. Ponadto przedstawiono analizę teoretyczna wykazującą że „Fraternal Dropout" ma działanie podobne do regularyzacji stabilizujących aktywacje w sieciach rekurencyjnych.

Sekcja 2.6 oraz praca „Improving the Performance of Neural Networks in Regression Tasks Using Drawering".

W pracy przedstawiono technikę regularyzacji sieci neuronowych rozwiązujących problem regresji przez uczenie ich na dodatkowym zadaniu klasyfikacji próbki do jednej z „poszufladkowanych" wartości. Wykazano że takie rozszerzenie sieci w trakcie uczenia poprawia jej zdolność generalizacji na dwóch zadaniach testowych. Wskazano również intuicje pomagające w doborze hiperparametru równoważącego podczas uczenia zadania regresji i klasyfikacji.

2.2 Uwagi polemiczne

W pracy Classifier-Agnostic Saliency Map Extraction ekstraktor mapy i klasyfikator są powiązane, choć dalsze eksperymenty wskazują, że wygenerowane maski skutecznie mylą też inne klasyfikatory. Intuicyjnie, musi tak być ponieważ wskazanie obiektu pierwszoplanowego jest zadaniem niezależnym od stosowanego klasyfikatora, i metoda CASME stanowi błyskotliwy przykład rozwiązania tego zadania. Co działby się jednak dla obrazów wieloznacznych, lub zawierających wiele obiektów? Który zostałby wskazany jako istotny? I czy w ogóle możemy mówić o CASME jako o metodzie tłumaczącej działanie klasyfikatora, skoro ignoruje ona całkowicie działanie klasyfikatorów innych, niż ten z którym współdzieli enkoder?

Bardzo chciałbym zobaczyć kontynuację prac dotyczących poprawę niezawodności klasyfikatorów przez zwiększenie ich selektywności, do czego pierwszym krokiem jest praca „Leakage-Robust Classifier". W szczególności chciałbym zobaczyć czy podobne techniki umiałyby wyjaśnić zjawisko „kruchych

cech" wskazane np. w „Adversarial Examples are not Bugs, they are Features"[2].

Prace Fraternal Dropout ma nieostatecznie opracowane state of te art: Fraternal Dropout stanowi przypadek szczególny zaproponowanych w 2014. roku pseudo-ensembles[3], ale niestety praca nie komentuje tej zależności. Brak tego odniesienia jest o tyle zaskakujący, że mechanizm pseudo-ensembles jest przedstawiony jako podstawa w podejściach z którymi bezpośrednio porównany był „Fraternal Dropout", np. sieciami Π Laina i Alia.

3 Konkluzja końcowa

Przedstawione prace, stanowiące jedynie fragment dorobku Doktoranta (jego profil Google Scholar wykazuje też inne prace powstałe w okresie jego studiów doktorskich) dobrze wskazują na wszechstronność i biegłość Autora w technikach uczenia maszynowego i uczenia głębokiego: Autor stosuje uczenie nadzorowane, uczenie ze wzmocnieniem oraz łączy metody nadzorowane z metodami generatywnymi. Podawane przez Autora intuicje uzasadniające dobór metod i hiperparametrów wskazują na zrozumienie przedmiotu.

Na uwagę zasługują licznie współprace naukowe nawiązane przez Doktoranta. Jego prace powstały w najlepszych światowych laboratoriach zajmujących się uczeniem głębokim.

Za najważniejszą pracę uważam „Classifier-Agnostic Saliency Map Extraction": z jednej strony wprowadza ona bardzo praktyczne narzędzie pozwalające na zrozumienie istotnych cech danych. Z drugiej strony, mimo prostoty pomysłu przyświecającego pracy, jej wykonanie wymagało kreatywnego połączenia wielu technik, wskazując tym samym na umiejętności implementacyjne Doktoranta.

Stwierdzam, że przedłożona mi do recenzji rozprawa p. mgra Konrada Żołny spełnia z nawiązką wymagania stawiane w Ustawie rozprawom doktorskim i wnoszę o jej dopuszczenie do publicznej obrony. Ze względu na wszechstronność podejmowanych tematów, oraz wprowadzenie przez Autora wielu użytecznych technik analizy i regularyzacji sieci wnoszę o wyróżnienie rozprawy doktorskiej.

dr hab. Jan Chorowski

[2] https://papers.nips.cc/paper/8307-adversarial-examples-are-not-bugs-they-are-features.pdf

[3] https://papers.nips.cc/paper/5487-learning-with-pseudo-ensembles.pdf