dr hab. Norbert Jankowski, prof. UMK

Katedra Informatyki Stosowanej, Uniwersytet Mikołaja Kopernika

ul. Grudziądzka 5, 87–100 Toruń

norbert@umk.pl, tel: (0 56) 6113307

Recenzja
rozprawy doktorskiej

„Safety and Trustworthiness of Deep Learning in Computer Vision – With Application of Out-of-Distribution Detection Techniques"

mgra Kamila Szyca

Pan mgr Kamil Szyc w swoim doktoracie zajmuje się analizą kwestii bezpieczeństwa użycia sieci neuronowych, a konkretnie głębokich sieci neuronowych, do rozpoznawania obrazów. Główną techniką jaką analizuje w doktoracie są algorytmy rozpoznające czy dany obiekt pochodzi z tego samego rozkładu danych co dane uczące głęboką sieć neuronową. To ważna kwestia ponieważ sieć neuronowa z natury rzeczy nie ma wbudowanej ochrony przed taką sytuacją a jak wiemy nauczona sieć neuronowa nie nadaje się do wiarygodnej klasyfikacji danych spoza rozkładu danych uczących. Wyobraźmy sobie, że sieć była uczona rozpoznawania 0 i 1, a po nauczaniu ktoś poprosi o klasyfikację obiektu pikseli, który będzie odpowiadał cyfrze 8. Bynajmniej nie ma co się spodziewać 8 jako wyniku klasyfikacji, a często wykorzystywany *soft-max* z przyjemnością zrobi z tego zero. Jednak bardziej kłopotliwe przypadki także były opisywane w literaturze, jak chociażby znamienny "one pixel attack", gdzie autorzy publikacji pokazali, że sieć neuronowa może być łatwo zmanipulowana. Takie przykłady i szereg innych w literaturze pokazują, że temat jest istotny, szczególnie gdy mamy do czynienia z wdrożeniami np. w medycynie obrazowej.

Rozdział drugi poświęcony jest wprowadzeniu w świat różnych modeli sieci neuronowych. Autor rozpoczął od podstawowego modelu sieci neuronowych i pierwszych artykułów w tym temacie i kontynuował prezentację poprzez najważniejsze osiągnięcia, czyli sieci konwolucyjne, sieci GAN, różne sieci rekuren-

cyjne w tym LSTM i ostatnio powszechnie używane transformery. Przy opisie modelu LSTM wzory są właściwe, ale rysunek wzięty z pewnej publikacji nie jest właściwy, tj nie obrazuje rekurencji h^{t-1} przy bramach i wejściu (lub rysunek jest uproszczony). Tutaj można było zacytować pierwszą pracę o LSTM S. Hochreitera i J. Schmidhubera.

Druga część rozdziału została poświęcona krótkiemu omówieniu najważniejszych odmian uczenia sieci neuronowych poprzez spadek gradientu. Wspomniano także o metodach regularyzacji spadku gradientu, a także technice dropout czy powiększaniu ilości danych poprzez ich transformacje. Omówiono także krótko najważniejsze funkcje aktywacji neuronów. Opisano również typy najczęściej wykorzystywanych warstw w sieciach CNN a następnie zaprezentowano sporo popularnych architektur głębokich sieci neuronowych.

W ostatnim podrozdziale autor omawia najważniejsze z punktu widzenia rozprawy doktorskiej publikacje, algorytmy i techniki związane z bezpieczeństwem sieci neuronowych. Omówione zostały różne podstawy problemów wiarygodności uczenia sieci neuronowych, które niemal zawsze są standardowo pomijane w pogoni za najlepszymi algorytmami uczenia. Przede wszystkim jednak opisano metody klasyfikacji danych spoza rozkładu danych treningowych. To właśnie analiza tych metod w następnym rozdziale stanowi kluczowe osiągnięcie niniejszej rozprawy doktorskiej. Autor wybrał i zaprezentował więc tutaj następujące algorytmy: MaxSoftMax, MaxLogits, Mahalanobis distance based methods, Local Outlier Factor (LOF), Open-Set Nearest Neighbor, MDistance, Extreme Value Machine (EVM), OpenMax i Odin. Metody te zostały opisane dość zwięźle. Zdarzało się jednak, że opisy były nieco zbyt skrótowe. Czasami można użyć większej ilości wzorów niż opisu językowego aby uzyskać większą precyzję, jak np. w metodzie LOF w kontekście wyznaczania wartości LDR, czy w metodach EVM i OpenMax. Należy podkreślić, że pośród opisanych metod metoda MDistance została zaproponowana przez doktoranta i podobnie, jak np. LOF, działa badając na swój sposób gęstość w wybranym punkcie klasyfikacji.

Na początku rozdziału trzeciego, gdzie autor definiuje jego główny cel i tym samym cel rozprawy, chyba doszło do pewnych chochlików językowych. Autor napisał: „the safety and trustworthiness of AI models can be assured only when the models are aware that the unknown data exist". Myślę, że świadomość to raczej zbyt silne wymaganie i raczej jej konieczności nie uda się udowodnić. Sądzę, że autor miał na myśli to, co napisał już w pierwszych rozdziałach rozprawy — że chce badać, czy dane pochodzą z tego samego rozkładu co dane treningowe. Ponieważ jest wręcz oczywiste, że po procesie uczenia modele sieci neuronowych są wykorzystywane do klasyfikacji prędzej nie znanych danych, które przecież mogą pochodzić z tego samego rozkładu danych.

Na zdecydowanie pozytywną ocenę zasługuje plan badań omawianego rozdziału. Autor precyzyjnie i właściwie opisał poszczególne cele do weryfikacji, a także sposoby ich uzyskania.

W teście z Tab. 3.2 mamy porównanie skuteczności metody MaxSoftmax i Mahalanobis, bazujące na efektach sieci CNN o różnej złożoności. Widać naprawdę dużą przewagę metody Mahalanobis. Autor dodatkowo pokazuje, że Mahalanobis działa oczekiwanie dobrze (w odróżnieniu od słabego MaxSoftmax) tylko przy rozpoznawaniu szumu, a gdy ma rozpoznawać elementy z SVHN czy CIFAR-100 tak dobrze już nie jest. Nie zawsze też większa złożoność sieci przekładała się na lepsze poprawności testów.

W teście z Tab. 3.3 mamy możliwość porównania wielu różnych metod OoD ze sobą podczas klasyfikacji przy użyciu sieci ResNet-101. Klasyfikacja, jak poprzednio, obejmuje elementy z części testowej CIFAR-10 od szumu, CIFAR-100 i SVHN. Tu także widać przewagę metody Mahalanobis nad wszystkimi innymi niezależnie od tego, czy w klasyfikacji mamy szum, SVHN czy CIFAR-100. Drugie w rankingu są metody LOF i LOF-D (czyli metoda LOF zmodyfikowana przez autora rozprawy). Te metody stabilnie plasują się tuż po metodzie Mahalanobis. To są konkluzje z pominięciem metod ODIN i Unified Framework z powodu ich dodatkowych założeń, co jak też pisze autor, może wpływać na wiarygodność rezultatów.

W Tab. 3.5 pokazano wyniki testu bazującego na obrazach o dużej rozdzielczości z wykorzystaniem różnych modeli CNN, m. in. AlexNet, Res-Net-18, VGG-16. Tutaj sytuacja się odwraca pomiędzy metodami Mahalanobis, LOF i LOF-D. Tym razem bardzo zdecydowanie wygrywa LOF-D. Inne metody spoza tej trójki często wręcz nie działają. Ciekawe dysproporcje można zaobserwować analizując różnice ze względu na typ sieci. Rezultaty dla AlexNet są przeciętne, a dla sieci takich jak WideResNet-101-2 wręcz bardzo dobre. To ewidentnie wynika z tego jaką jakość generalizacji uzyskały w procesie uczenia. Ten test pokazuje bardzo ciekawe cechy metod OoD i ich zależności od typu sieci CNN, a także odmienności rozkładu testowego.

Kolejny test dotyczy możliwości metody Mahalanobis w obliczu wielowymiarowych danych. Autor zaproponował test, w którym dane są generowane sztucznie za pomocą chmur punktów o odpowiednio zdefiniowanych wartościach średnich i wariancjach. Jednak, moim zdaniem, nie można tak łatwo obserwacji z tego testu przenieść na analizę danych zebranych z sieci CNN, ponieważ sieci neuronowe tylko po procesie inicjalizacji wag będą porównywalne co do rozkładów. Po procesie nauczania dzięki uzyskaniu dyskryminacji wartości wag są dalekie od początkowego rozkładu. Chyba właśnie dlatego w poprzednich testach widzieliśmy pozytywny efekt korzystania z metody Mahalanobis w kontekście klasyfikacji szumu jako OoD.

W teście, którego wyniki są przedstawione w Tab. 3.6 widzimy, że w różnych konfiguracjach sieci i danych uczących każda z metod potrafi w niektórych okolicznościach być ewidentnie gorsza. Różnice, które obserwujemy rzeczywiście potrafią być istotne. Należy jednak zwrócić uwagę, że przy małej ilości zbiorów danych i małej ilości konfiguracji sieci to nie jest bardzo zaskakujące i potwierdza to co już prędzej widzieliśmy, że w niektórych sytuacjach najlepsza była metoda Mahalanobis a kiedy indziej ewidentnie LOF-D zaproponowana przez doktoranta. Jednakże niezaprzeczalnie mamy do czynienia z pewnym rodzajem niestabilności.

Następna analiza dotyczy modelu Unified Framework (UF). Metoda UF konsoliduje kilka elementów: komitet cech, transformacja wejścia i korzystanie z Metody Mahalanobis. Autor przeprowadził ciekawe badanie, które wyznacza, co rzeczywiście składa się na sukces tej metody. Dlatego zaproponował aby zrobić szereg testów na bazie UF z włączonym lub wyłączonym komitetem cech i transformacją wejścia. To daje oczywiście 4 testy, z których łatwo wywnioskować, że oba elementy są znacząco potrzebne do uzyskania najciekawszych rezultatów.

Kolejnym krokiem była propozycja autora, aby zastąpić użycie Mahalanobis'a poprzez LOF i LOF-D (por. Fig. 3.7 i Tab. 3.7). Przeprowadzony test pokazał, iż był to bardzo słuszny krok (AUC >80% z 50%). Autor podkreśla słusznie niedogodność metody UF, polegającą na konieczności podglądania części danych spoza rozkładu danych uczących, aby wyznaczyć wagi komitetu UF. Oczywiście taki tuning można widzieć jako "uproszczoną dyskryminację" co musi wpłynąć pozytywnie na rozpoznawanie danych spoza rozkładu. Jednak nie zawsze jest to niewykonalne (zależy od aplikacji).

Porównane zostały metoda EVM i jej modyfikacje zaproponowane przez autora oraz metoda LOF. Zaproponowane zmiany okazały się korzystne, ale nie są jednak lepsze, niż wcześniej omawiana metoda LOF.

Kolejny podrozdział poświęcony jest badaniom metod ekstrakcji cech z sieci CNN do algorytmów OoD. Oprócz GAP zbadano rezultaty dla GMP, GAP-All, CroW i SCADA. Autor zauważa, że zmiany są zdecydowanie pozytywne. Warto zauważyć dodatkowo, że szczególnie CroW daje ogromny wzrost DTACC, ale dla zbiorów o rozkładzie znacząco różnym od rozkładu danych treningowych, czyli np. dla szumu, ImageNet-O czy Places365. Natomiast nieco gorsze rezultaty są dla zbioru CIFAR-100. To bardzo ciekawe, bo wzmocnienie dla znacząco różnych rozkładów jest bardzo pozytywne, ale widać jeszcze jakiś brak informacji *do jakich obiektów są te cechy używane*, czyli cechy są, ale jakby brakowało informacji, do czego te cechy są.

Kolejne badanie miało na celu analizę wpływu redukcji wymiarowości poprzez PCA. W procedurze tej

najpierw następuje normalizacja danych potem wybór cech z PCA i znów normalizacja. Nie do końca rozumiem po co robić normalizację po PCA. Być może rezultaty mogłyby być nieco lepsze gdyby zrezygnować z normalizacji po PCA. Analiza wyników pokazała, że trzeba właściwie wybrać ilość składników głównych, aby nieco polepszyć rezultaty dla metod LOF*. W przypadku innych algorytmów OoD raczej obserwujemy spadki jakości.

Przebadano także wpływ wielu różnych metod powiększania zbioru danych na działanie algorytmów OoD. Część metod powiększania zbiorów dała bardzo pozytywny efekt, np. CropAndPad czy MixUp. Może i efekt był spodziewany w ogólności, ale szczegóły pokazujące które metody na ile powiększają już wcale oczywiste nie są. Dlatego test można uznać za zdecydowanie interesujący.

Absolutnie bardzo ciekawy jest kolejny test, który bada stabilność a raczej niestabilność procesu uczenia głębokich sieci. Różnice jakie powstały w rozróżnianiu danych spoza rozkładu pod wpływem dwóch trochę różnych procesów uczenia ale prowadzących to modeli nauczonych na tym samym poziomie poprawności robią wrażenie. Choć nie ukrywam, że też bym się tego spodziewał. Dodam, iż chętnie bym widział aby każdy z testów A i B powtórzyć choć z 5 lub 10 razy (aby różnice leżały tylko w losowej inicjalizacji!), aby popatrzeć na średnie poprawności i ich odchylenia. Może da się to zobaczyć na obronie?

W ostatnim teście autor pokazuje skuteczność rozpoznawania ataków na sieć ResNet-101. Porównane zostały różne typy ataków z (ciekawszymi) różnymi typami algorytmów OoD. Najlepsze algorytmy OoD w większości mają poprawność w okolicy 80% ale są wyjątki, jak zawsze. Ciekawy okazał się tutaj algorytm MDistance, które poprzednio był w cieniu innych a w wykrywaniu ataków radził sobie zdecydowanie.

Podsumowanie

Autor doktoratu wykazał się bardzo szeroką wiedzą z zakresu sieci neuronowych, ze szczególnym uwzględnieniem najnowszych osiągnięć w tej dziedzinie. Poznał też wiele algorytmów OoD. Zaproponował szereg ulepszeń istniejących algorytmów, które okazały się zdecydowanie ciekawe. Zaproponował też nowe algorytmy OoD. Doktorant zaproponował bardzo wiele ciekawych analiz porównawczych, które nie tylko weryfikowały stawiane hipotezy, ale także naprowadzały na kolejne ciekawe badania. Wszystko to omówione zostało szczegółowiej w powyższej części recenzji.

Na uwagę zasługuje starannie dobrana i bardzo bogata bibliografia (niemal 300 pozycji). Praca charakteryzuje się też dużą poprawnością naukowo-techniczną.

Warto też wspomnieć, że Pan Kamil Szyc jest autorem 11 publikacji z czego trzy są wysoko punktowane.

Kończąc oceniam bardzo pozytywnie dorobek doktoranta. Dlatego uważam, że zaprezentowana rozprawa spełnia warunki dotyczące prac doktorskich i stawiam wniosek o dopuszczenie jej autora do dalszych etapów przewodu doktorskiego.

Toruń, 10.09.2022

Norbert Jankowski