WROCŁAW UNIVERSITY OF SCIENCE AND TECHNOLOGY

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

By

KAMIL SZYC

DOCTORAL THESIS

Supervisor:
Henryk Maciejewski
PhD, DSc, Assoc. Prof.

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY
DEPARTMENT OF COMPUTER ENGINEERING

06.06.2022 r.
Szyc

Wrocław, V 2022

Abstract

Although the newest computer vision models achieved impressive accuracy, further challenges in the problem of image classification still exist. Due to the broad applications of these algorithms in real-life, improving security and trustworthiness seems to become the most important task nowadays for researchers.

There are many challenges and threats to deep learning approaches connected with the above. One of them is susceptibility to natural and adversarial attacks. The network can be easily fooled with special kinds of images. The goal of adversarial attacks lies in preparing new images by adding additional unique noise that forces the network to point out some class with high certainty despite the image not presenting that class. Images that do not belong to any known classes - however, the network returns high certainty results for them, can be considered as natural attacks. Another challenge is the robustness of the network. The robustness is the ability of the network to classify similar (but not within the same distributions) images to the training examples – for instance, working with extra distortions like with rotated images, with different weather conditions (e.g., fog, rains), or obscured by some object. The robustness is also the ability to return similar outputs for similar inputs, increasing the stability of the response. It connects with the next challenge - interpretability. Modern models work like "black boxes", so it is impossible to interpret the factors influencing the outputs. However, we - people - would like to understand why the network responds in that way to get to know the principles.

All of the above can be connected to one of the most important threats, in our opinion – the close-set nature of classification. The artificial network usually is forced to choose one winning class, while there is no "I do not know" or "unknown" class. Often the classifier returns a high certainty answer to a random label (from a human perspective). When we will "open" the model by adding the ability to detect unknown data distribution, we increase the safety of the ML model.

The main thesis of this dissertation is that the safety and trustworthiness of AI models can be assured only when the models can distinguish between known and unknown data - therefore, we showed and thoroughly researched the Out-of-Distribution detection techniques. Still, there are no clear conclusions and recommendations in the literature on which methods are most successful for a given recognition problem. Furthermore, our research suggests that the best method depends on pairs of known and unknown samples. However, the OoD samples are unknown by their nature. In this thesis, we looked at OoD from a practical perspective. The safety and trustworthiness models must work as open-set classification – we focused on how to force classic close-set models to be OoD aware. Our contributions are as follows.

We analyzed the selected popular OoD methods. We showed there is no best OoD approach, and it should depend on the tested ID and OoD pair. We showed limitations and analyzed the assumptions of the Extreme Value Machine (EVM) algorithm and instability of parametric models based on MultiVariate Normal (MVN) distribution, i.e., using the Mahalanobis distance, in high dimensional data. We showed that the nonparametric, density-based LOF approach performs better than based on MVN or logits in many cases. We see a lack of comparison between LOF and other benchmark methods in the literature.

We showed the efficiency of the OoD method based on large-scale (e.g., the high image resolutions or the high number of known classes) dataset benchmarks. Many OoD methods

are evaluated only on a low-scale dataset, which is insufficient in real-life problems. We conclude that the OoD methods based on logits work poorly, and Mahalanobis is slightly worse compared to other methods on a large-scale. The LOF seems to be the most stable method.

We noticed the significant influence of the feature extraction strategy on improving the efficiency of OoD detection. The standard method uses only Global Average Pooling (GAP). However, various approaches can focus on different components (e.g., on edges, patterns, or whole objects), so for different pairs of ID and OoD, different feature extractors can be useful in separating data. We recommend it as a new hyperparameter. We showed that reducing the size of feature vectors leads to severe efficiency deterioration for many methods. The LOF-based methods seem to be the only method we recommend using together with dimensional reduction.

We showed that OoD detection can filter many adversarial examples. Moreover, we recommend choosing a different feature extraction strategy than GAP (notice that the GAP is the base for the classifier) - it may improve the efficiency of detecting attack samples.

We showed the importance of proper data augmentation techniques in OoD detection problems and robustness.

We researched the problems that occur with unknown examples of detection. We analyzed the problem with instability and experimented with repeatability of the OoD detection methods – and concluded that results in literature should be taken with caution. The slightly different model state can change the OoD method's efficiency drastically. Moreover, we showed the mismatch between the image and feature space - i.e., similar images (in our human understanding) can generate distant features, and images from different classes can be close to each other in feature space. We conclude that the common usage of near and far OoD examples definitions is inaccurate.

KEYWORDS

Image Classification – Computer Vision – Out-of-Distribution Detection – Open-Set Classification – Deep Learning – Robustness – Security and Trustworthiness of AI – Features Extraction Methods – Adversarial Attacks – Convolutional Neural Networks

Streszczenie (Abstract - Polish Version)

Najnowsze modele wizyjne osiągają imponującą dokładność, jednak wciąż istnieją kolejne wyzwania w problemie klasyfikacji obrazów. Szerokie zastosowanie takich modeli w praktycznych projektach sprawia, że obecnie poprawa bezpieczeństwa i wiarygodności sztucznej inteligencji wydaje się najważniejszym zadaniem dla badaczy.

Z powyższym zagadnieniem wiąże się również wiele innych wyzwań i zagrożeń. Jednym z nich jest podatność na ataki naturalne i typu adversarial (z ang. adversarial attacks). Sieć można łatwo oszukać za pomocą specjalnych obrazów. Celem ataków typu adversarial jest przygotowanie nowych obrazów poprzez dodanie dodatkowego unikalnego szumu, który zmusza sieć do wskazania wybranej klasy z dużą pewnością, mimo że obraz nie przedstawia tej klasy. Obrazy, które nie należą do żadnych znanych klas, a mimo to sieć jest pewna swojej błędnej odpowiedzi nazywamy atakami naturalnymi. Kolejnym wyzwaniem jest krzepkość (z ang. robustness) sieci. Krzepkość to zdolność sieci do klasyfikowania obrazów podobnych do przykładów treningowych - na przykład obrazów, które zostały w pewien sposób zniekształcone. Przykładem takich zniekształceń mogą być obrócone zdjęcia, zdjęcia zrobione w różnych warunkach pogodowych (np. we mgle lub w deszczu) lub gdy główny obiekt jest zasłonięty. Krzepkość to także zdolność do zwracania podobnych wyników dla podobnych danych wejściowych, określana także jako stabilność odpowiedzi. Łączy się to z kolejnym wyzwaniem - interpretowalnością. Współczesne modele działają jak "czarne skrzynki", więc nie da się zrozumieć czynników wpływających na wyniki. My - ludzie - chcielibyśmy jednak zrozumieć, dlaczego sieć reaguje w ten sposób, aby poznać zasady jej działania.

Wszystko to można powiązać z jednym z najważniejszych, moim zdaniem, zagrożeń – klasyfikacją w zbiorze zamkniętym. Sztuczna sieć jest zazwyczaj zmuszona do wyboru jednej zwycięskiej klasy, podczas gdy nie ma klasy "nie wiem" lub "nieznane". Często klasyfikator zwraca odpowiedź o wysokim stopniu pewności do losowej (z ludzkiego punktu widzenia) etykiety. Gdy "otworzymy" model, dodając możliwość wykrywania nieznanego rozkładu danych, zwiększymy bezpieczeństwo modelu ML.

Główną tezą tej rozprawy jest to, że bezpieczeństwo i wiarygodność modeli sztucznej inteligencji można zapewnić tylko wtedy, gdy modele te będą umieć rozróżnić dane znane od nieznanych - dlatego też pokazałem i dokładnie zbadałem techniki wykrywania danych nieznanych (z ang. Out-of-Distribution, OoD detection). Wciąż jednak w literaturze nie ma jednoznacznych wniosków i zaleceń dotyczących tego, które metody są najskuteczniejsze. Co więcej, przeprowadzone w tej pracy badania sugerują, że najlepsza metoda zależy od pary danych znanych i nieznanych. Jednakże dane nieznane są z natury nieokreślone, co sprawia, że problem jest trudny. W tej pracy skupiono się na praktycznym podejściu do klasyfikacji w zbiorze otwartym. Bezpieczne i wiarygodne modele muszą działać jako klasyfikatory otwarte - skupiłem się na tym, jak zmusić klasyczne modele, aby były świadome danych nieznanych. Poniżej opisany jest najważniejszy wkład tej pracy.

Dokładnie przeanalizowałem wybrane metody wykrywania danych nieznanych. Wykazałem, że nie ma najlepszej metody, a jej wybór powinien zależeć od badanej pary znanych i nieznanych danych. Znalazłem ograniczenia i zbadałem założenia metody EVM oraz niestabilność modeli parametrycznych opartych na wielowymiarowym rozkładzie normalnym (z ang. MultiVariate Normal distribution, MVN) tj. wykorzystujących odległość Mahalanobisa. Nieparametryczna metoda LOF oparta na gęstości w wielu przypadkach

sprawdza się lepiej, niż metoda oparta na MVN lub bazująca na odpowiedzi z sieci (z ang. logits). W literaturze niestety brakuje porównań metod typu LOF z innymi.

Sprawdzono skuteczność metod wykrywania danych nieznanych w dużej skali (np. obrazy z dużą rozdzielczością czy bazy danych z dużą liczbą klas znanych). Wiele metod jest ocenianych jedynie na zbiorach danych w małej skali, co jest niewystarczające w przypadku rzeczywistych problemów. Pokazałem, że metody oparte na odpowiedzi z sieci działają słabo, a metoda Mahalanobisa nie działa tak dobrze. LOF wydaje się być najbardziej stabilną metodą.

Zauważyłem znaczący wpływ metody ekstrakcji cech na poprawę skuteczności wykrywania danych nieznanych. Klasycznie wykorzystuje się jedynie metodę GAP (z ang. Global Average Pooling). Jednak inne podejścia mogą skupiać się na innych elementach obrazu (np. na krawędziach, wzorach lub całych obiektach), więc dla różnych par danych znanych i nieznanych, różne ekstraktory cech mogą być przydatne w separacji danych znanych i nieznanych. Zaproponowałem dobór ekstrakcji cech jako nowy hiperparametr. Ponadto sprawdziłem wpływ redukcji rozmiaru wektorów cech. Metody oparte na LOF wydają się być jedynymi metodami, przy których taka redukcja ma sens.

Pokazałem, że przy użyciu metod służących do wykrywania danych nieznanych można odfiltrować wiele przykładów ataków typu adversarial. Zalecam wybór innej strategii ekstrakcji cech niż GAP (cechy GAP są wykorzystane przez klasyfikator) - może to poprawić skuteczność wykrywania.

Wykazałem, że dobór odpowiedniej techniki rozszerzania danych (z ang. data augmentation) wpływa znacząco na efektywność metod wykrywa danych nieznanych i krzepkość sieci.

Zbadałem również problemy występujące przy wykrywaniu danych nieznanych. Przede wszystkim zbadałem problem niestabilności omawianych metod - doszedłem do wniosku, że wyniki podane w literaturze należy traktować z ostrożnością. Nieco inny stan modelu może drastycznie zmienić skuteczność metody. Ponadto wykazałem możliwą rozbieżność między przestrzenią obrazów a przestrzenią cech - tj. podobne obrazy (w naszym ludzkim rozumieniu) mogą generować odległe cechy, a całkowicie różne obrazy mogą być bliskie sobie w przestrzeni cech. Powszechnie stosowane definicje przykładów danych nieznanych bliskich (z ang. near OoD) i dalekich (z ang. far OoD) są nieprecyzyjne.

Słowa Kluczowe

Klasyfikacja obrazów – Widzenie komputerowe – Wykrywanie danych nieznanych – Klasyfikacja w gupie otwartej – Głębokie uczenie – Bezpieczeństwo i wiarygodność sztucznej inteligencji – Metody ekstrakcji cech – Ataki typu adversarial – Konwolucyjne sieci neuronowe