Barbara Laskowska*

Automatyczne rozpoznawanie treści nielegalnych filmów typu CSAM za pomocą klasyfikatora częściowo splatającego kolejne klatki materiału wideo[1]

Streszczenie

Artykuł zawiera opis jednej z metod automatycznego wykrywania treści nielegalnych typu CSAM, którą przetestowano podczas badań przeprowadzanych w projekcie APAKT. Zaproponowane rozwiązanie wykorzystuje klasyfikator Temporal Shift Module (TSM), model sieci głębokiej do wydajnego rozpoznawania aktywności na plikach wideo. Zastosowano metodę z transferem wiedzy, żeby stosunkowo niedużą liczbą danych uczących nauczyć model skutecznego rozpoznawania treści pornograficznych i nielegalnych na filmach. Przeprowadzono testy skuteczności klasyfikacji na danych neutralnych legalnej i nielegalnej pornografii. W artykule wskazano również związane z tym tematem badawczym problemy, które wynikają z charakterystyki danych. Ponadto zwrócono uwagę na konieczność dalszych prac nad zapewnianiem bezpieczeństwa dzieci w cyberprzestrzeni.

Słowa kluczowe: bezpieczeństwo cyberprzestrzeni, analiza wideo, uczenie maszynowe, sieci neuronowe, sieci głębokie, transfer learning, Child Sexual Abuse Material, CSAM

* Mgr inż. Barbara Laskowska, Zespół Złożonych Systemów, Instytut Automatyki i Informatyki Stosowanej, Wydział Elektroniki i Technik Informacyjnych, Politechnika Warszawska, e-mail: barbara.laskowska.dokt@pw.edu.pl, ORCID: 0000-0003-0958-9969.

1 Praca finansowana przez Narodowe Centrum Badań i Rozwoju w ramach projektu CYBERSECIDENT/455132/III/NCBR/2020.

Wstęp

W otwartej sieci, jaką jest internet, łatwo udostępniać wszelkie treści zarówno sprawdzone, jak i fikcyjne. Dane, wyniki badań, informacje o sobie, pomysły, przemyślenia, zdjęcia, filmy, nagrania audio przechowywane na różnych serwerach są dostępne z prawie każdego miejsca na świecie. Nie wszystkie materiały są dopuszczalne przez prawo, a po wykryciu naruszeń zachodzi konieczność zablokowania takich treści i często ścigania autora lub osobę, która je upowszechnia, oraz wszczęcia postępowania karnego. Do treści ściganych z urzędu zaliczają się materiały CSAM, tj. przedstawiające wykorzystywanie seksualne dzieci (Child Sexual Abuse Material) w formie zdjęć i nagrań wideo.

Dyżurnet.pl[2] to specjalny zespół utworzony do wykrywania tego typu nadużyć i reagowania na nie. Praca w nim jest trudna i obciążająca psychicznie. Żeby wspomóc pracę moderatorów, w zespole powstał projekt APAKT[3], którego głównym zadaniem było utworzenie systemu współpracujących ze sobą klasyfikatorów służących do rozróżniania treści nielegalnych, zawierających pornografię z udziałem osób nieletnich na zdjęciach i filmach umieszczanych w internecie.

Wykrywanie nielegalnych treści

Do klasyfikacji zdjęć i wideo są stosowane głębokie sieci neuronowe[4] składające się z kilkudziesięciu warstw, a każda z nich zawiera setki lub nawet tysiące parametrów. Żeby dobrze nauczyć, czyli wytrenować, tak dużą sieć i ustalić wartości milionów parametrów, potrzeba ogromnej liczby przykładowych danych uczących. O ile można zebrać dużą liczbę neutralnych filmów, o tyle tych zawierających treści nielegalne nie ma dość dużo na potrzeby typowego

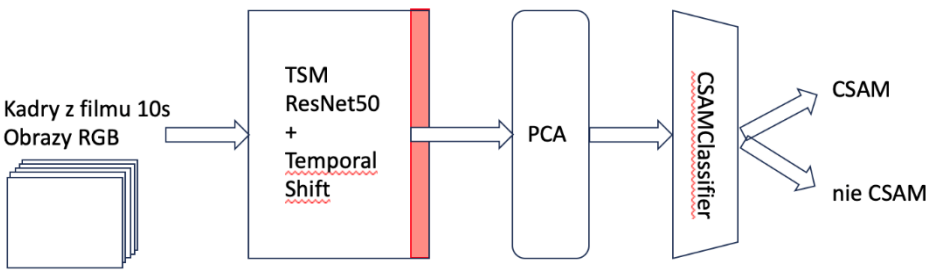
2 Dyzurnet.pl – zespół Naukowej i Akademickiej Sieci Komputerowej. NASK-PIB został wskazany jako jeden z Zespołów Reagowania na Incydenty Komputerowe zgodnie z ustawą o krajowym systemie cyberbezpieczeństwa.

3 Projekt APAKT – system reagujący na zagrożenia bezpieczeństwa dzieci w cyberprzestrzeni ze szczególnym uwzględnieniem pornografii dziecięcej; finansowany przez Narodowe Centrum Badań i Rozwoju w ramach projektu CYBERSECIDENT/455132/III/NCBR/2020.

4 Uczenie głębokie (deep learning) – podkategoria uczenia maszynowego polegająca na tworzeniu głębokich sieci neuronowych składających się z wielu warstw neuronów do rozpoznawania zdjęć, wideo, przetwarzania języka naturalnego.

zadania uczenia modelu. Ponadto są one dostępne w oryginale jedynie dla pracowników zespołu Dyżurnet.pl i niemożliwy jest ani wgląd w ich naturę, ani spreparowanie dodatkowych treści o takim charakterze na potrzeby opracowania dobrego modelu klasyfikującego. Żeby system mógł działać z odpowiednią skutecznością, postanowiliśmy użyć podejścia z transferem wiedzy (transfer learning) polegającego na wykorzystaniu gotowych dużych modeli sieci neuronowych wytrenowanych do rozwiązywania ogólnego problemu klasyfikacji treści na ogromnych zbiorach danych. Modele te są następnie douczane na docelowych danych, których jest znacznie mniej. Wykorzystaliśmy również zbiory danych z legalną pornografią, żeby nauczyć model rozpoznawania filmów o charakterze pornograficznym bez uwzględniania wieku osób biorących udział w nagraniu.

Do analizy materiału wideo wykorzystaliśmy m.in. model TSM (Temporal Shift Module)[5] służący do klasyfikacji czynności wykonywanej na filmie. Model został nauczony na bazie Kinetics 400[6] zawierającej ponad 280 tys. filmów z 400 kategoriami czynności o charakterze powszechnym i neutralnym, takimi, jak: gotowanie jajka, tańczenie zumby, jedzenie spaghetti. Model TSM jest oparty na dwuwymiarowej sieci ResNet50[7] służącej do analizy obrazów. Sieć została następnie wzbogacona o moduł, który przesuwa dane pomiędzy kolejnymi kadrami. Przesunięcie nie jest kosztowne obliczeniowo, a umożliwia analizę wideo porównywalną do modeli trójwymiarowych sieci splotowych.



Źródło: Opracowanie własne.

Schemat przetwarzania wideo

5 J. Lin, C. Gan, S. Han, *TSM: Temporal Shift Module for Efficient Video Understanding*, 2018, https://doi.org/10.48550/arXiv.1811.08383 [dostęp: 15.06.2023].

6 W. Kay i in., *The Kinetics Human Action Video Dataset*, 2017, https://doi.org/10.48550/arXiv.1705.06950 [dostęp: 15.06.2023].

7 B. Koonce, *ResNet 50* [w:] *Convolutional Neural Networks with Swift for Tensorflow*, Berkeley, CA 2021.

Wejściem do modelu są obrazy RGB, które są kadrami z filmu długości maks. 10 sekund. Model przetwarza obrazy powielone na różnych filtrach, wyznacza ostatecznie wektor 16 384 cech filmu. Następnie na podstawie tych cech klasyfikuje wynik do jednej z 400 kategorii.

Zadanie klasyfikacji pornografii dziecięcej ograniczyliśmy w naszej adaptacji modelu TSM do rozpoznawania dwóch klas – CSAM oraz nie CSAM. Do nauki końcowego klasyfikatora mamy mało danych uczących. Chcieliśmy zredukować liczbę parametrów w końcowym klasyfikatorze stosownie do wielkości zbioru uczącego. W tym celu zastosowaliśmy kompresję PCA redukującą 16 384 pierwotnych cech filmu do 200 najistotniejszych. Tak przygotowane dane są wprowadzane na wejście do końcowego klasyfikatora CSAM Classifier, który jest siecią neuronową z jedną warstwą ukrytą.

Przedstawione podejście pozwala zachować zalety dużej sieci nauczonej przetwarzania danych wideo oraz wyspecjalizować mały klasyfikator do rozpoznawania specyficznych danych.

Przeprowadzone testy i ich wyniki

W celu sprawdzenia poprawności powyższego rozwiązania przeprowadziliśmy kilka eksperymentów. Pierwszym etapem było wypróbowanie rozwiązania do zadania rozróżniania legalnej pornografii od filmów neutralnych. Następnie sprawdziliśmy klasyfikację filmów CSAM i neutralnych oraz rozróżnianie CSAM od legalnej pornografii. Ostatnim zadaniem była klasyfikacja wieloklasowa, w której dane dotyczyły wszystkich trzech kategorii.

Dane uczące pochodziły z trzech źródeł. Filmy neutralne to podzbiór filmów bazy Kinetics 400. Filmy legalnej pornografii pochodziły z bazy „brazylijskiej" oraz z serwisu pornhub.pl, pornografii dziecięcej z zasobów zespołu Dyżurnet.pl.

Jedynie specjalny zespół może mieć dostęp do danych nielegalnych, więc ich przetwarzanie mogło odbywać się tylko w odpowiednio zabezpieczonym środowisku obliczeniowym udostępniającym wyłącznie abstrakcyjne wysokopoziomowe wektory cech filmów.

Uśrednione wyniki przeprowadzonych badań są zawarte w tabeli 1. Najlepsze wyniki osiągnęliśmy dla filmów pornograficznych i neutralnych. Rozróżnienie filmów pornografii legalnej od nielegalnej okazało się dużo trudniejszym zadaniem, gdzie osiągnęliśmy dokładność na poziomie 70%. Trudność klasyfikacji filmów o charakterze pornograficznym jest widoczna również w wynikach klasyfikacji wieloklasowej.

Tabela 1. Uśrednione wyniki przeprowadzonych badań klasyfikatorów treści wideo

Zadanie	Dokładność (%)	Czułość – CSAM (%)	Czułość – filmy pornograficzne (%)	Czułość – filmy neutralne (%)
Porn/neutral	95,71	-	97,74	93,88
CSAM/neutral	95,22	98,36	-	92,31
CSAM/porn	69,57	72,40	66,90	-
CSAM/porn/neutral	75,78	70,00	69,42	88,85

Źródło: Opracowanie własne.

Problemy w klasyfikacji materiału CSAM

Zadanie wykrywania treści nielegalnych jest zadaniem szczególnym z kilku powodów. Przede wszystkim naukowcy przygotowujący modele sztucznej inteligencji nie mogą zobaczyć danych, na których pracują. Jest to sytuacja niespotykana i jednocześnie utrudniająca badania. Trudno określić przyczyny niepowodzeń klasyfikacji, jeżeli nie wiadomo, na jakich danych uczył się model, jaka była ich charakterystyka. Można jedynie się domyślać, jakie możliwe modyfikacje algorytmu mogłyby poprawić wynik, a potem sprawdzać je w praktyce. Takie podejście jest trudne, czasochłonne i niejednokrotnie niedające oczekiwanych wyników.

Filmy zabezpieczone w zasobach Dyżurnet.pl mają różny format, różną długość. Nie wiadomo, którą część kadru w materiale oryginalnym zajmuje osoba nieletnia, ani w którym fragmencie filmu się pojawia. Film oznaczony jako nielegalny może zawierać tylko kilkusekundową scenę faktycznie w kategorii CSAM. Taka charakterystyka danych utrudnia wyselekcjonowanie dobrych przykładów uczących dla klasyfikatorów.

Kategoria CSAM nie jest kategorią jednorodną. Treści te mogą się od siebie znacząco różnić w zależności od tego, do której podkategorii należą. Wyznaczone przez Dyżurnet.pl kategorie CSAM: 1) małoletni i dorośli – występuje czynność seksualna dziecka oraz jest obecna osoba dorosła; 2) tylko małoletni – czynność seksualna dziecka lub nastolatka bez udziału i obecności dorosłych; 3) w obecności małoletniego – aktywność seksualna osób dorosłych wykonywana w obecności dziecka, lecz bez jego udziału; 4) fokus – brak czynności seksualnej dziecka, zbliżenie na obszar genitalny lub analny; 5) pozowanie seksualne – brak czynności seksualnej, eksponowany obszar analny lub genitalny dziecka.

Dodatkowa kategoria treści podobnych do powyższych, a jednak legalnych, to child nudity przedstawiająca nagość dzieci bez zabarwienia erotycznego oraz child erotism, która przedstawia dzieci pozujące już w kontekście erotycznym, ale jeszcze niepornograficznym.

Rozbicie problemu klasyfikacji na wszystkie te kategorie z osobna z jednej strony pozwala na dokładniejszą specjalizację sieci, ponieważ dane wewnątrz kategorii są do siebie bardziej podobne. Z drugiej strony, wymaga większej liczby danych uczących, żeby dla każdej kategorii można było zdefiniować reprezentatywną próbkę danych.

Zakończenie

Ochrona bezpieczeństwa dzieci i ich prawa do zdrowego rozwoju to jedno z podstawowych zadań społeczeństwa[8]. Działania chroniące dzieci w cyberprzestrzeni nie ograniczają się do edukacji w sposobie korzystania z nowoczesnych technologii i uwrażliwienia dorosłych na nieodpowiednie treści. Powstaje bezpieczny ekosystem internetowy, w których treści są dostosowane do wieku i kontrolowane przez narzędzia kontroli rodzicielskiej. Do ochrony dzieci i ich wizerunku w cyberprzestrzeni zaliczamy również zwalczanie przestępstw seksualnych poprzez szybkie wykrywanie treści przedstawiających takie czyny. Internet daje dzieciom dużo możliwości edukacji, rozwoju kompetencji społecznych czy rozrywki, ale również stanowi zagrożenie.

System APAKT jest odpowiedzią na jedno z zagrożeń. Będzie wspierał ludzi codziennie reagujących na treści w internecie w wykrywaniu materiałów nielegalnych i szkodliwych. Konieczna jest ciągła praca nad systemem zarówno w jego utrzymaniu, jak i dalszym rozwoju, żeby skutecznie spełniał swoje zadanie.

System można rozwijać wielotorowo. Należy zweryfikować działanie wdrożonego systemu i sprawdzić, które moduły wymagają dopracowania. Równoczesne prace nad odpowiednim oznaczaniem danych uczących będą wspierały poprawę działania algorytmów.

8 K. Badźmirowska-Masłowska, *Child protection in cyberspace*, „Cybersecurity and Law" 2019, nr 1, s. 213–224.

Opisany w tym artykule algorytm osiąga dobre wyniki w zadaniu rozróżniania filmów pornograficznych od neutralnych. W dalszych badaniach należy sprawdzić, czy połączenie tego modelu z modelem klasyfikującym wiek osób na filmie poprawi klasyfikację treści CSAM.

Bibliografia

Aggarwal Ch.C., *Neural Networks and Deep Learning*, Cham 2018.

Badźmirowska-Masłowska K., *Child protection in cyberspace*, „Cybersecurity and Law" 2019, nr 1.

Lin J., Gan C., Han S., *TSM: Temporal Shift Module for Efficient Video Understanding*, 2018, https://doi.org/10.48550/arXiv.1811.08383 [dostęp: 15.06.2023].

Kay W. i in., *The Kinetics Human Action Video Dataset*, 2017, https://doi.org/10.48550/arXiv.1705.06950 [dostęp: 15.06.2023].

Koonce B., *ResNet 50* [w:]: *Convolutional Neural Networks with Swift for Tensorflow*, Berkeley, CA 2021.

Automatic recognition of the content of CSAM using a classifier that shifts part of the channels from successive frames of video

Abstract

The paper describes one of the methods of automatic recognition of CSAM materials, which was tested during the research under the APAKT project. The proposed solution is based on Temporal Shift Module (TSM), a model of a deep neural network created for efficient human activities rocognition in video. We applied transfer learning method for training the model with a relatively small number of training data to succesfully rocognize films with pornografic and illegal content. We conducted some tests of classification of films from three categories: neutral films, legal pornography and illegal pornografic videos (CSAM). In this paper we present problems that are connected with this research topic that come from the characteristic of the data. We also show that further works are needed to keep children safe in cyberspace.

Key words: cybersecurity, computer video analysis, machine learning, neural networks, deep neural networks, transfer learning, Child Sexual Abuse Material, CSAM