УДК 004.932
DOI: 10.18523/2617-3808.2022.5.68-71

Бучко О. А., Нгуєн С. Б. В.

КЛАСИФІКАЦІЯ КОНФІДЕНЦІЙНИХ ЗОБРАЖЕНЬ ІЗ ВИКОРИСТАННЯМ НЕЙРОННОГО ХЕШУ

Запропоновано підхід із застосуванням нейронного хешу для розв'язання задачі класифікації конфіденційних зображень. Головна ідея алгоритму полягає у пошуку схожих зображень – таких, що слугуватимуть взірцем для визначення класів. Алгоритм використовує хеш-коди, що дає змогу забезпечити приватність світлин користувачів. Розглянуто псевдоадаптивність мережі.

Ключові слова: нейронні мережі, хешування, найближчі сусіди, навчання на прикладах.

Вступ

Швидкість технічного прогресу вплинула на те, що людство, як ніколи раніше, генерує величезну кількість інформації, використовуючи свої персональні пристрої — смартфони, ноутбуки, планшети. Зображення завантажуються на безліч різноманітних платформ, як-от соціальні мережі, месенджери, вебсервіси та інші застосунки, що своєю чергою становить велику небезпеку для людини та її персональної інформації. Конфіденційність користувача довгий час експлуатують у мережі Інтернет: використовуючи такі прості речі, як вік, вага, національність, релігія, вподобання тощо, зацікавлені сторони заманюють потенційного клієнта у пастку пропозицій і послуг. Конфіденційна інформація, яка може міститися в особистих зображеннях, інколи не розпізнається їх користувачами як небезпечна для розголошення, а тому може легко бути поширена в мережі самим власником без роздумів. Варто розглянути й іншу сторону таких дій, оскільки не тільки користувача треба захищати від такої інформації, а й інших від неї. Нерідко такий контент може бути чутливим для інших, особливо для деяких категорій користувачів. Тому для розв'язання цієї проблеми використаємо підхід нейронного хешу й дослідимо ефективність цього алгоритму для домену конфіденційних зображень. Певні досягнення у цій сфері має компанія Apple, яка досліджувала нейронний хеш для розпізнавання CSAM (англ. child sexual abuse material) [1].

Опис алгоритму

Алгоритм поєднує у собі декілька методів і задач, які є фундаментальними. Перелік кроків алгоритму такий:

1) трансферне навчання;
2) вилучення ознак;
3) хешування ознак (local sensitivity hashing);
4) пошук найближчих сусідів;
5) визначення ймовірності класів.

Головними компонентами алгоритму є нейронна мережа, яка генерує вектори вилучених ознак для зображень, і проіндексований набір зображень (хеш-таблиці), в яких зберігаються знання про певний домен, наприклад конфіденційні зображення. Отримані за допомогою нейронних мереж хеш-коди називатимемо нейронним хешем.

Головна ідея алгоритму передбачає колізії хеш-кодів для зображень, які є схожими, за рахунок подібності їхніх векторів вилучених ознак. Такий підхід називають чутливим хешуванням (англ. local sensitivity hashing) [6]. Отримані хеш-коди можуть бути ідентичними або відрізнятися на певну відстань за Гемінгом. Для збільшення повноти або влучності результатів можна використати декілька хеш-таблиць, які використовуватимуть різні хеш-функції, а результати будуть об'єднуватися або перетинатися відповідно.

Класифікація класів відбувається за рахунок пошуку схожих зображень-зразків (пошук найближчих сусідів), які зберігаються в хеш-таблиці у вигляді нейронних хеш-кодів. Імовірність кожного класу визначається як відношення зображень-зразків цього класу до найбільшого класу в запиті. Належність до класу передбачає, що зображення має ймовірність, яка є більшою за певний поріг, що може визначатися окремо для кожного класу.

Алгоритм досліджено на основі попередньо натренованої нейронної мережі ResNet50, яка має велику кількість шарів, що дає змогу робити

точніші передбачення, ніж аналогічні мережі [2]. Використовують набір даних «The Visual Privacy (VISPR) Dataset», який містить 68 різних класів [7]. Прикладами класів є «номерний знак», «кредитна карта», «паспорт», «водійське посвідчення», «студентський квиток», «підпис», «повна голота», «релігія» тощо. Кількість використаних зображень з вибраного набору даних становить 22 200 зображень. Програмну реалізацію може бути виконано за допомогою будь-яких сучасних програмних інструментів, наприклад бібліотеки «fast.ai» [5], яку використано в цьому дослідженні.

Аналіз та оцінка алгоритму

Для оцінки роботи алгоритму використано набір із 2200 зображень. Оцінку роботи алгоритму проведено на основі відомих метрик, що застосовують для класифікації зображень:

– точність: accuracy = (true positive + true negative) / (total sample size), частка правильно прогнозованих позитивних і негативних результатів;
– повнота або чутливість: recall = true positive / (true positive + false negative), частка загального числа позитивних зразків, яку було знайдено;
– влучність: precision = true positive / (true positive + false positive), прогностичні значущості позитивного й негативного результатів;
– F_1-score: визначається як гармонійне середнє значення влучності та повноти [8], тобто

$$F_1 = 2\frac{precision * recall}{precision + recall}.$$

Результати роботи алгоритму для різної кількості проіндексованих зображень-зразків наведено у табл. 1.

Таблиця 1

Оцінка роботи алгоритму до вилучення класів

Кількість зразків	accuracy	recall	precision	F_1-score
1000	0,24	0,14	0,25	0,18
5000	0,41	0,30	0,45	0,36
10 000	0,52	0,46	0,56	0,50
20 000	0,54	0,48	0,57	0,57

На основі отриманих результатів можна виявити, що влучність роботи алгоритму є більшою, ніж її повнота, що свідчить про високу чутливість алгоритму й допустимість хибно-негативних результатів. Така характеристика не є бажаною, оскільки за таких умов алгоритм пропускатиме конфіденційні зображення як безпечні. Загальна оцінка алгоритму також досить низька – F_1-score становить 0,57.

Така оцінка пов'язана з таксономією обраного набору зображень, оскільки він містить багато класів з абстрактними і суперечливими визначеннями, наприклад «релігія» або «національність». Тому, щоб нівелювати вплив таких класів, створено спеціальний «білий» список класів, які мають чітке представлення. До цього списку було включено класи: «автомобільний номер», «кредитна картка», «паспорт», «пошта», «чек», «квитки», «підпис», «імейл». Оцінку роботи наведено в табл. 2.

Таблиця 2

Оцінка роботи алгоритму для обраних класів

Кількість зразків	accuracy	recall	precision	F_1-score
1000	0,36	0,21	0,40	0,28
5000	0,63	0,45	0,65	0,53
10 000	0,86	0,78	0,84	0,81
20 000	0,91	0,85	0,89	0,87

За рахунок вилучення абстрактних класів було досягнуто кращої оцінки алгоритму, а саме 0,87 для F_1-score. Для покращення цих метрик знадобиться більший набір зображень-зразків, які зможуть передбачити більше випадків для кожного класу.

Псевдоадаптивність

Головною перевагою алгоритму є псевдоадаптивність, тобто можливість продовжувати збільшувати кількість класів і випадків для класів, не змінюючи ваги нейронної мережі, як у традиційному адаптивному навчанні [4]. Така властивість дає змогу реалізувати зворотний зв'язок для користувачів, які зможуть доповнювати й додавати нові класи власноруч. Алгоритм можна віднести до алгоритмів, які навчаються на невеликій кількості прикладів [3].

Щоб дослідити псевдоадаптивність, використано зображення котів, що є зовсім іншим доменом. Для експерименту використано 1000 зображень, які було додано як зразки до сформованих хеш-таблиць. Щоб оцінити роботу, було зібрано набір із 50 зображень котів і 50 зображень собак. Ознайомитися з метриками можна у табл. 3.

Таблиця 3

Оцінка псевдоадаптивності алгоритму

Кількість зразків	accuracy	recall	precision	F_1-score
1000	0,86	0,80	0,91	0,85

Отримані результати свідчать про спроможність системи продовжувати вчитися на основі прикладів навіть з інших доменів. Однак залишається проблема низької повноти результатів, що може бути критичною для деяких предметних галузей.

Псевдоадаптивність може вирішити питання великої різноманітності екземплярів, наприклад студентських квитків, які мають різний форм-фактор і вигляд у різних країнах. Через зворотний зв'язок можна додавати зразки, які важко передбачити експерту, або зібрати приклади через їх конфіденційність.

Висновки

У процесі виконаної роботи було досліджено можливість використання нейронних хеш-кодів для класифікації зображень із конфіденційним вмістом. Такий підхід дає змогу зберігати конфіденційність персональних зображень користувачів за рахунок представлення зображень як хеш-коду. Проаналізовані метрики свідчать про високу залежність алгоритму від правильно побудованої таксономії і зображень-зразків. На основі таких метрик, як влучність і повнота, можна зробити висновок, що алгоритм гарно підходить для предметних галузей, які допускають більше хибно-негативних результатів, а отже для систем, для яких важлива релевантність результатів.

На основі цього алгоритму досліджено псевдоадаптивність алгоритму, що дає можливість збільшувати кількість класів і зразків і допускає зворотний зв'язок від користувачів. Зазначена властивість є головною перевагою алгоритму, що може стати головним критерієм для вирішення подібних проблем в інших предметних сферах.

Список літератури

1. Apple. CSAM detection: technical summary [Electronic resource] / Apple. – [S. l. : s. n.]. – 12 p. – Mode of access: https://www.apple.com/child-safety/pdf/CSAM_Detection_Technical_Summary.pdf.
2. Deep residual learning for image recognition [Electronic resource] / Kaiming He [et al.] // Microsoft research. – P. 12. – Mode of access: https://doi.org/10.48550/arXiv.1512.03385.
3. Generalizing from a Few Examples [Electronic resource] / Yaqing Wang [et al.] // ACM computing surveys. – 2020. – Vol. 53, no. 3. – 34 p. – Mode of access: https://doi.org/10.1145/3386252.
4. Goodfellow I. Deep learning (adaptive computation and machine learning series) / Ian Goodfellow, Yoshua Bengio, Aaron Courville. – [S. l.] : MIT Press, 2017. – 800 p.
5. Howard J. Deep Learning for Coders with Fastai and Pytorch: AI Applications Without a PhD / Jeremy Howard, Sylvain Gugger. – [S. l.] : O'Reilly Media, Incorporated, 2020. – 624 p.
6. Lee K. M. Locality-Sensitive hashing techniques for nearest neighbor search [Electronic resource] / Keon Myung Lee // International journal of fuzzy logic and intelligent systems. – 2012. – Vol. 12, no. 4. – Pp. 300–307. – Mode of access: https://doi.org/10.5391/ijfis.2012.12.4.300.
7. Orekondy T. Towards a visual privacy advisor: understanding and predicting privacy risks in images [Electronic resource] / Tribhuvanesh Orekondy, Bernt Schiele, Mario Fritz // 2017 IEEE international conference on computer vision (ICCV), Venice, 22–29 October 2017. – [S. l], 2017. – Mode of access: https://doi.org/10.1109/iccv.2017.398.
8. Sasaki Y. The truth of the F-measure [Electronic resource] / Yutaka Sasaki // Research fellow. – 5 p. – Mode of access: https://www.toyota-ti.ac.jp/Lab/Denshi/COIN/people/yutaka.sasaki/F-measure-YS-26Oct07.pdf.

References

Apple. (2021). CSAM detection: Technical summary. https://www.apple.com/child-safety/pdf/CSAM_Detection_Technical_Summary.pdf.

Goodfellow, I., Bengio, Y., & Courville, A. (2017). Deep learning (adaptive computation and machine learning series). MIT Press.

He, K., Zhang, X., Ren, S., & Sun, J. Deep residual learning for image recognition. Microsoft Research, 12. https://doi.org/10.48550/arXiv.1512.03385

Howard, J., & Gugger, S. (2020). Deep learning for coders with fastai and pytorch: AI applications without a phd. OReilly Media, Incorporated.

Lee, K. M. (2012). Locality-Sensitive hashing techniques for nearest neighbor search. *International Journal of Fuzzy Logic and Intelligent Systems*, *12* (4), 300–307. https://doi.org/10.5391/ijfis.2012.12.4.300

Orekondy, T., Schiele, B., & Fritz, M. (2017). Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *2017 IEEE international conference on computer vision (ICCV)*. https://doi.org/10.1109/iccv.2017.398

Sasaki, Y. The truth of the F-measure. *Research Fellow*, 5. https://www.toyota-ti.ac.jp/Lab/Denshi/COIN/people/yutaka.sasaki/F-measure-YS-26Oct07.pdf.

Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a Few Examples. *ACM Computing Surveys*, *53* (3), 1–34.

O. Buchko, S. B. Nhuien

CLASSIFICATION OF CONFIDENTIAL IMAGES USING NEURAL HASH

Humanity generates considerable information using its devices — smartphones, laptops, and tablets. Users upload images to different platforms, such as social networks, messengers, web services and other

applications, which greatly endanger their personal information. User privacy has been exploited on the Internet for a long time. Interested parties lure potential customers into a trap of offers and services using such information as age, weight, nationality, religion and preferences. The sensitive information that may be contained in personal images is sometimes not recognized by their users as dangerous to share and, therefore, can easily be shared online by the owner without a second thought.

This article inspects a neural hash algorithm for solving image classification tasks of confidential information and evaluates it via basic metrics. The main idea of the algorithm is to find similar images that will serve as an example for defining classes. The algorithm uses hash codes, ensuring users' privacy. The evaluation of the algorithm is based on "The Visual Privacy (VISPR) Dataset". The main components of the algorithm are a neural network that generates vectors of extracted features for images and an indexed set of images (hash tables) that store knowledge about a particular domain.

The critical aspect of the algorithm involves collisions of hash codes for similar images due to the similarity of their vectors of extracted features. The resulting hash codes can be identical or differ by a specific value of Hamming distance. Multiple hash tables with different hash functions are used to increase the recall or precision of the results. The effect of imperfect taxonomy was analyzed, which led to further filtration of abstract classes and increasing overall scores.

Also, the article investigates the "pseudo-adaptivity" of the algorithm - the ability to classify new classes and add new cases to existing classes that were not included in the training stages. Such ability may be crucial for domains with many image instances or classes.

Keywords: neural networks, hashing, nearest neighbor, instance-based learning.

Матеріал надійшов 14.08.2022