Multi-task learning model for detecting and filtering internet violent images for children

Le Kim Hoang Trung[1], Nguyen Van Thanh Vinh[1], Phan Le Viet Hung[1], and Nguyen Huu Nhat Minh[1]
[1] The University of Danang, Vietnam - Korea University of Information and Communication Technology, Danang, Vietnam

Correspondence: Nguyen Huu Nhat Minh, nhnminh@vku.udn.vn
Received: 01/05/2024, revised: 01/05/2024, accepted: 01/05/2024
Digital Object Identifier:: 0.32913/mic-ict-research-vn.v2023.n1.1212

Abstract: The Internet has emerged as an essential daily information access, but exposing children to inappropriate content can impair their early development. Existing content filtering methods exhibit limitations in accurately and efficiently detecting diverse inappropriate internet content. In this paper, we propose a multi-task learning model for detecting and filtering violent images to provide safer online experiences. The multi-task model is developed from the pre-trained lightweight base model such as MobileNetv2 to enable proper integration within web browser extensions. Pure training to detect violent images could raise false alarms in the classification results when the landscape or object images don't contain any human, hence we develop two joint learning tasks such as detecting humans and detecting violent images simultaneously. Our experiments demonstrate that the proposed multi-task approach with binary rule achieves 98.5% accuracy, outperforming the single-task model for detecting violent images by a margin. Thereafter, the multi-task model is also integrated into the web extension to detect and filter out violent images to prevent children from harmful content.

Keywords: *Multi-task learning, Violence detection, Web extension.*

Tiêu đề: **Mô hình đa tác vụ nhận diện và lọc hình ảnh bảo lực cho trẻ em**

Tóm tắt: Internet đã trở thành một phương tiện truy cập thông tin thiết yếu hàng ngày, nhưng việc trẻ em tiếp xúc với nội dung không phù hợp có thể gây hại cho sự phát triển ban đầu. Các phương pháp lọc nội dung hiện có gặp hạn chế trong việc phát hiện chính xác và hiệu quả các nội dung không phù hợp đa dạng trên Internet. Trong bài báo này, chúng tôi đề xuất một mô hình đa tác vụ để phát hiện và lọc hình ảnh bạo lực nhằm cung cấp trải nghiệm trực tuyến an toàn hơn. Mô hình đa nhiệm này được phát triển từ mô hình nền tảng nhẹ đã được huấn luyện trước như MobileNetv2 để có thể tích hợp hợp lý vào các tiện ích mở rộng của trình duyệt web. Việc huấn luyện để phát hiện hình ảnh bạo lực có thể gây ra cảnh báo sai trong kết quả phân loại khi các hình ảnh phong cảnh hoặc vật thể không chứa bất kỳ con người nào, do đó chúng tôi phát triển hai nhiệm vụ học liên kết như phát hiện con người và phát hiện hình ảnh bạo lực đồng thời. Kết quả thí nghiệm cho thấy đề xuất mô hình đa tác vụ với quy tắc nhị phân đạt độ chính xác 98.5%, vượt trội so với mô hình đơn tác vụ trong việc phát hiện hình ảnh bạo lực. Sau đó, mô hình này cũng được tích hợp vào tiện ích mở rộng của trình duyệt để phát hiện và lọc bỏ hình ảnh bạo lực nhằm ngăn chặn trẻ em tiếp xúc với nội dung có hại.

Từ khóa: *Học đa tác vụ, nhận diện bạo lực, tiện ích mở rộng Web*

I. INTRODUCTION

As technology advances rapidly, Internet access plays an increasingly vital role in human life. For children, the Internet provides vast information access including educational, entertainment, and social communication resources. However, along with these benefits, there is a worrying proliferation of violent images circulating uncontrollably, posing a threat to children. Exposure to such content can negatively impact the early development of children. Hence, there is an urgent need for effective measures to protect children from exposure to inappropriate images on the Internet.

Statistics show that 71% of children aged 15-17 in developed countries have accessed websites with violent images [1]. Exposure to negative images and videos can lead to severe consequences such as depression, increased bullying tendencies, and potentially irreversible damage to brain structure and cognition capabilities, especially

given their developing brains. Despite concerted efforts by governmental and non-governmental entities to address this predicament, existing solutions still lack the capacity for automated detection and filtering of risky content. Due to these technical gaps, children still readily access images and videos that are violent. Given these disturbing vulnerabilities, this research puts forward an AI system capable of identifying and obscuring negative elements in visual media. Integrating the model into web browsers would obstruct the proliferation of harmful materials, thereby safeguarding future generations from the detrimental influences of the online world. The proposed approach leverages intelligent algorithms to construct a bulwark against age-inappropriate content, aiming to uphold the safety of children during their digital encounters. As they traverse the largely unregulated online landscape, young users require robust protection from materials that can severely impact their developmental trajectory. By preemptively barring such exposures, this system endeavors to supplement ongoing initiatives focused on fostering a benevolent internet environment tailored uniquely to children's needs.

To address this challenge, this research has endeavored to build a deep learning model that has the ability to detect feature extraction from violent images, with a focus on safeguarding the online experience for children. By leveraging technological advancements and drawing insights from existing research, this research paper aims to contribute to ongoing efforts to create a safe, healthy environment for children from violence at an early age. Through meticulous research and development, this study aims to enhance our understanding of methods for filtering inappropriate images and provide practical solutions to mitigate risks associated with children's exposure to harmful online content. Additionally, the learning model will be integrated into a web extension capable of effectively recognizing and filtering inappropriate images for children is intended to promote positive online experiences for future generations.

There have been efforts and solutions to limit the prevalence of violent and negative images and video content online, but identifying all such images is a difficult problem to solve. New images can be continuously posted every second, and it's impossible to know if they are safe for children. This paper aims to address the following questions such as:

- How to develop a deep learning-based image classification model with high accuracy in detecting violent images?

- If integrated into a browser extension, can the model automatically blur/block inappropriate images in real time?

Objectives and contributors:

The main objective of this research is to develop a multi-task learning model capable of classifying and identifying violent images on the Internet. The proposed model employs the pre-trained MobileNetv2 model as the shared layers, trained on two learning tasks with specific-task layers for detecting human as well as violent images. The learning model increases recognition capabilities and will be integrated into web browser extensions to automatically blur or block inappropriate images from children by training and collection of Existing image data from the internet.

In evaluating the effectiveness of the solution, the research will be contributed by the process of collecting image data on violent images and labeling the dataset. Additionally, the design and training of the neural network model through multiple times to improve performance. Moreover, the development of the extension and evaluation through multiple testing experiments also contributes significantly to the research process.

This initiative aspires to set new standards for protecting children from the hazards latent within the internet's vast unsupervised terrain. The envisioned automation of visible content moderation through an AI-based approach aims to alleviate the damages wrought by inadequate safeguards while upholding youthful innocence in the digital age.

II. RELATED WORKS

In this section, we provide an overview of existing studies and research related to developing systems for detecting and filtering violent images on the Internet, particularly in relation to children. This section explores the approaches, techniques, challenges, and limitations identified by previous research. Additionally, it serves as a basis for analyzing these studies and how they inform the methodology of the current paper.

In existing research, there were different approaches to this topic. Research by David Alexander on the use and effectiveness of Artificial Intelligence on internet websites shows the future possibilities and promising results in applying the many potentials of Artificial Intelligence [2]. In addition, there is research on the use of AI in work using violent images to threaten children and also focuses on techniques using Deep learning in image recognition on social networking platforms [3]. Identifying violent images by dividing them into multiple cases the identification of violent images provides a higher level of accuracy and coverage than the research of Muhammad Ramzan [4].

The recent methods for detecting violent images include feature-based methods and neural network-based methods. Feature-based methods [5] focus on identifying specific image points or possibly sounds within images/videos related to violence. On the other hand, neural network-based methods concentrate on building and developing flexible

neural networks to understand the characteristic relationships within image data. Neural network-based methods generally yield better results compared to feature-based methods, but it is worth noting that they require a large amount of complex and accurate training data.

We also have approached many software available on the software market and researched the methods they have used such as NetNanny [6], CyberSister[7], MaxProtect[8], etc. most of which are still limited in flexibility, active and accessible web extension browsers. In the existing paper, we explore and investigate various methodologies of inappropriate image detection to gauge aspects of viability and application in real-world contexts. Specifically, we analyze methodologies including URL Blacklists/Whitelists. In URL lists maintaining dynamically updated forbidden URL datastores via web crawlers enables preemptive filtering by browser cross-referencing requested pages against known banned domains. Benefits include real-time detection and easy integration into existing platforms [9].

Image recognition algorithms are also used with array models instead of pixels, they give better results than modern algorithms that model pixels, whereas the adult Web page bag recognition is carried out using multi-instance learning based on the combination of classifying texts, images, and videos in Web pages, Both the speed and the accuracy for recognizing the Web adult content are increased, in contrast, to detect Web pages one-by-one and also called Hybrid Text/ Metadata Heuristic, in every content supplementary signals like text captions, webpage markup, and origin domains assist violent/explicit detection when used adjunctively with computer vision. This fusion approach counterbalances blind spots between modalities yet requires more complex orchestration [10]. There are also content filtering and detection techniques based on Chroma to detect image content, so this approach follows classic machine learning methods including RandomForest, NeuralNetwork. With "Adult & Safe" images classified and accuracy up to 88% [11]

Challenges and Limitations of Existing Methods:

We are always aware that detecting sensitive and violent content on the internet is very important to protect users, especially children, from harmful experiences online. However, current methods face many challenges and limitations that affect the research and development process. The first challenge encountered is the complexity of images and video content, the constant and diverse appearance of training data, and the ability to accurately identify all cases of violent images due to their complexity and diversity. For example, the image of two people rubbing each other's heads also makes it difficult for the model to know whether it is an act of love or violence.

The next challenge is that if you want the model to be accurate then the work that comes with it. It is the training data that must be suitable for each classification and requires a large amount of data, which causes difficulty in training and evaluating the model and slows it down significantly. The next step is to apply them to the web extension. This must ensure the model is small enough for processing speed to be almost real-time, ensuring accurate model recognition along with high processing speed. Therefore, the biggest challenge for this research is the proposed model for web extensions must be lightweight, high accuracy, and suitable for almost all devices.

III. DATASET CONSTRUCTION

1. Public Datasets

The "Public Datasets" section of this paper incorporates several publicly available datasets essential for training and evaluating human activity and violence detection systems. We utilized two public datasets as follows:

- **Real Life Violence Situations Dataset [12]**
  This dataset, introduced by Soliman et al., comprises 1000 violent and 1000 non-violent videos sourced from YouTube. The violent videos encompass real street fight scenarios in diverse environments and conditions, while the non-violence videos depict various human actions such as sports, eating, and walking. We extracted five random frames from those videos, however, this is limited by a number of scenarios and is not enough to produce a high-performance detection model in diverse practical scenarios.
- **Human Action Recognition (HAR) Dataset [13]**
  The HAR Dataset, available on Kaggle, provides a collection of $12,000$ images depicting various human actions. However, we have only selected $6,000$ images. This dataset is added to provide more images for normal situations without violence.

These datasets play a pivotal role in the development and assessment of the proposed system, ensuring its efficacy in detecting and filtering violent images on the internet, particularly with a focus on children. Therefore, we collectively construct a preliminary dataset including $10,000$ samples for violent and $3,000$ for normal human action.

2. Manual Collection

While public datasets served as a valuable starting point, we recognized that they had significant limitations in diversity and coverage of realistic scenarios. Labeled images from public datasets, though useful, did not fully capture the wide variability of violent images propagated through

social media. To complement the public data, we undertook a manual collection process focused on gathering samples from major social platforms. This allowed us to obtain more varied and representative examples reflective of actual online content.

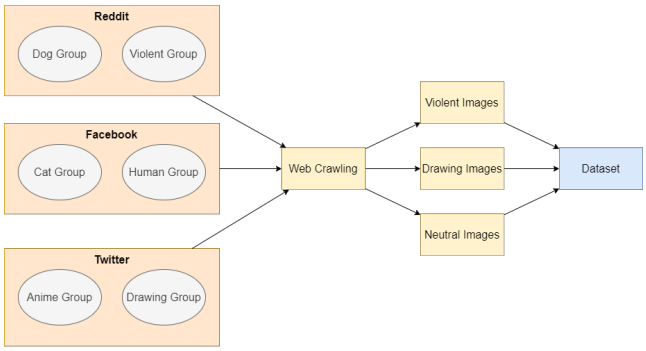

Figure 1. How web crawling work

We manually crawled additional samples from three public social media platforms such as Facebook, Reddit, and Twitter that had violent images or others as shown in Fig.1. We accessed existing online social network groups to collect more samples in a variety of realistic images from many kinds of groups. For instance, to collect more violent images we searched for the group titled "Fight Situations" and downloaded all shared images within those groups for around 2 months. All images from the "Fight Situations" Facebook group were annotated with the label “Violent". On the other hand, the normal cases include groups containing animals, nature, anime, and drawing pictures. As a result, we manually added $5,000$ normal, $2,000$ violent into our dataset containing $23,000$ images. While automatically crawling social media enabled the collection of a large volume of labeled image data, however, we detected that it contained significant noise in automatic labeling. For example, the "Fight Situations" group likely contained some nonviolent images that would be incorrectly labeled as “Violent”.

Ultimately, we accepted this trade-off because social media provided access to an exceptionally large and diverse range of image data. Our strategy for dataset collection made the learning model more robust compared to training only on limited public datasets. Hence, diverse data sources enabled us to develop a deep learning model with superior generalization ability for detecting violent images in practical internet images.

IV. PROPOSED METHOD

1. Multi-task learning approach

Multi-task learning [14] trains a model to perform multiple related learning tasks simultaneously. This technique provides several advantages over specialized single-task models. By sharing representations between learning tasks, the model learns generalized features that benefit all of the tasks. Training signals from the multiple objectives also regularize the model and prevent overfitting. Additionally, multi-task models require fewer parameters than multiple independent single-task models, improving efficiency.

To enhance our detection of violent images, we employ a multi-task learning approach for detecting human appearance as well as violence simultaneously. The learning model architecture shares base layers such as convolutional layers in MobileNet [14] that extract generalized visual features for both learning tasks. In that way, detecting human appearance as a learning task helps to improve violence detection in two aspects. Firstly, it increases the generalization of the detection model focus on human actions. Secondly, the output from human appearance detection helps to correct misclassified normal pictures due to our limited datasets. Furthermore, the proposed multi-task learning approach leverages the MobileNetv2 model [15], which has been pre-trained on the ImageNet dataset as the transfer learning approach[15]. The pre-trained MobileNetv2 was chosen due to the lightweight design that makes it well-suited for integration into web extensions for existing browsers (e.g., Chrome, Edge, Brave) to provide real-time filtering on the client side.

The model will be continued training on our collected dataset from public datasets and manual collection. After passing through the MobileNet layers for feature extraction, the features will be fed into separate task-specific layers for each class (violence, human). These layers will produce predictions on whether the image has violence or human figures as shown in Fig.2. The reason we included the human appearance detection task is that it provides additional features related to the presence of humans in the images, which aids the model in detecting violent actions involving humans. Also, the output from human appearance detection helps to correct the wrong prediction from violent detection learning tasks. This often happens in realistic unseen natural scenes or normal animal pictures that do not exist in the training datasets.

Specifically, the predictions from the two output layers will be combined to determine the final action of blurring or not blurring the image. The predictions will be combined and represented as a binary string “ab":

- “a" indicates violence, “0" if the image is violent, else “1"
- “b" indicates human presence, “0" if the image contain human, else “1"

Based on the predefined **binary rule**, images predicted as “11" will be blurred, while images predicted as “00", “01",

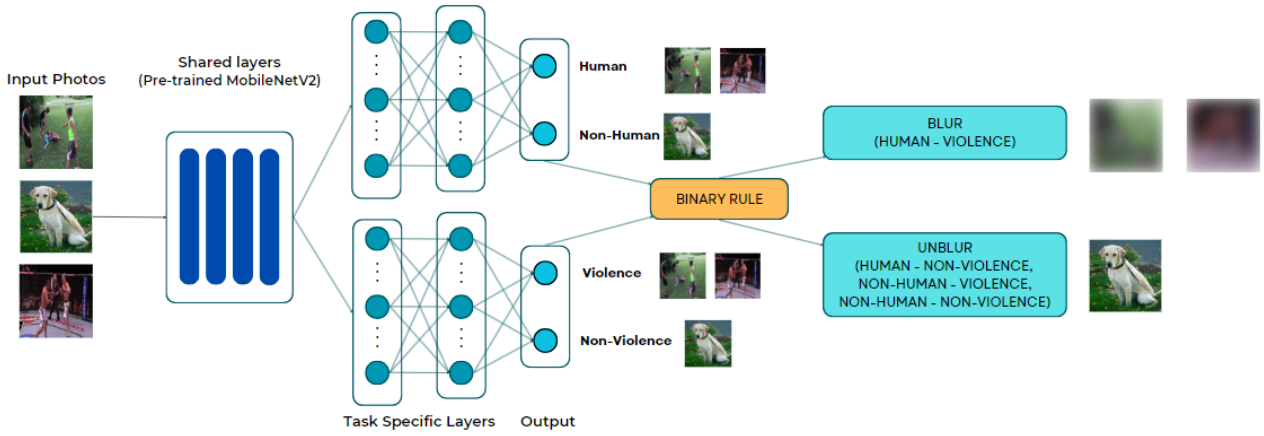

Figure 2. The proposed multi-task learning approach.

or “10" will remain unblurred. This multi-task approach allows the model to leverage inter-related signals between detecting violence, and humans in images to improve the performance of identifying images that should be blurred. Thereby, the proposed method ultimately boosts the robustness of the learning model as shown in our evaluation section.

V. RESULTS

We evaluated the performance of our proposed multi-task learning model on a dataset of 23,000 images. The dataset consisted of 10,000 images for violent or non-violent detection and 13,000 images for detecting human appearance. The non-human images include a variety of objects (such as drawings, animals, flowers, etc.).

In training our multi-task learning model, we utilized the **categorical cross-entropy** loss function for both the human detection and violence detection tasks. This choice of loss function was crucial in optimizing the model’s performance across both tasks. The categorical cross-entropy effectively measures the divergence between the predicted probabilities and the true class labels, ensuring that the model learns to distinguish accurately between categories. By summing this loss function from each task, we ensured consistent and reliable training outcomes, which contributed to the high accuracy levels observed in our evaluations.

$$\mathcal{L}_{CE} = -\sum_{i=1}^{classes} y_i \log(\hat{y}_i) \quad (1)$$

where $\hat{y}$ is the predicted result and y_i is the true label.

The dataset was splitted into 70/30 for training/testing data. For the violence detection task, the proposed multi-task model achieved an accuracy of **97.6%** while the human detection task obtained an accuracy of **96%** as shown in Table 1. To evaluate the performance of the multi-task model against single-task models, we trained separate models for violence detection using the same dataset.

The single-task violence detection model achieved an accuracy of 95.2% when using the additional collected data while obtaining 94% with two public datasets. Hence, this result demonstrates that the multi-task learning approach, by sharing representations between related tasks, improves performance over single-task models by 2.4%. Thereby, we adopted the binary rule to enhance the robustness of the proposed approach and obtain the highest accuracy performance i.e., 98.5%. Even though we did not count exactly in practice, the **binary rule** exhibits consistently improving the robustness of detecting realistic images on social networks.

To conclude, we showed that using multi-task learning to predict the relation between human presence and violent images enhanced the accuracy of detecting violent images, unlike models that use only one task. The model achieved promising results on the test set, indicating its potential for integration into web browsers as a client-side filtering extension.

VI. WEB EXTENSION INTEGRATION

To enable real-time filtering of violent images during web browsing, we developed a web extension integrating our trained multi-task learning model. The extension is implemented for all browsers with Chromium core by using Chrome Extensions API [16].

The extension loads the file of our model locally within the web browser for inference. For every web page that users visit, the extension will scan and multi-task learning model evaluates each appeared image and decides whether to blur it or not. If the decision is “blur", the extension dynamically modifies the image tagged as inappropriate by applying a

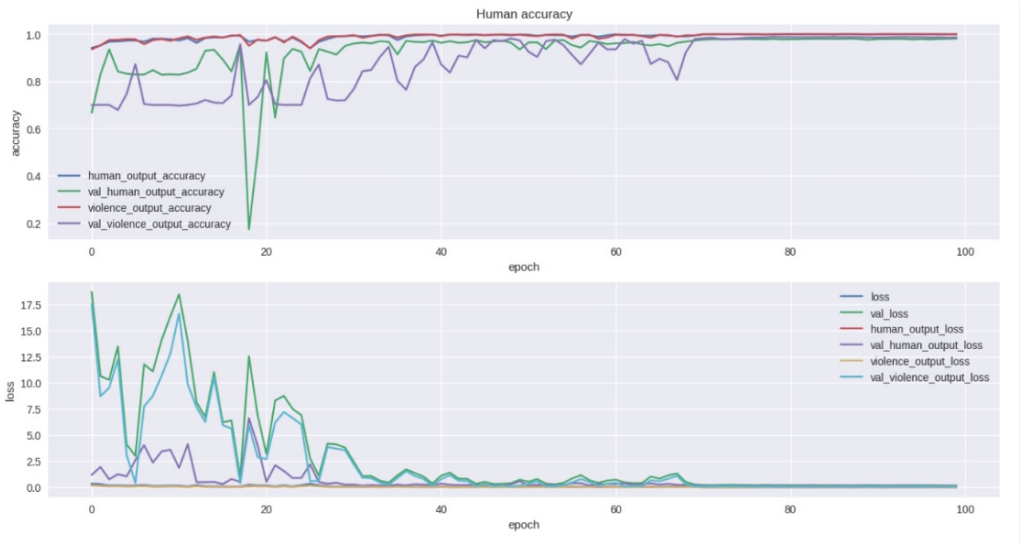

Figure 3. The accuracy performance and loss of training and validation for human appearance detection task and violent detection task.

Learning Task	Single-Task	Single-Task-PD	Multi-Task	Multi-Task-B
Violence Detection	95.2%	94%	97.6%	**98.5%**
Human Detection	-	-	**96.4%**	**96.4%**

Table I
TESTING ACCURACY OF LEARNING TASKS USING MULTI-TASK LEARNING AND SINGLE-TASK LEARNING APPROACHES. **SINGLE-TASK-PD** IS THE SINGLE-TASK MODEL USING ONLY TWO PUBLIC DATASETS WITHOUT ADDITIONAL DATA. AND THE PROPOSED **MULTI-TASK-B** APPROACH USES THE MULTI-TASK MODEL WITH THE ADDITIONAL BINARY RULE.

blur effect, else does nothing. This requires fast interaction to guarantee a low delay for the users. Only those visible images to the users will be detected and blurred iteratively.

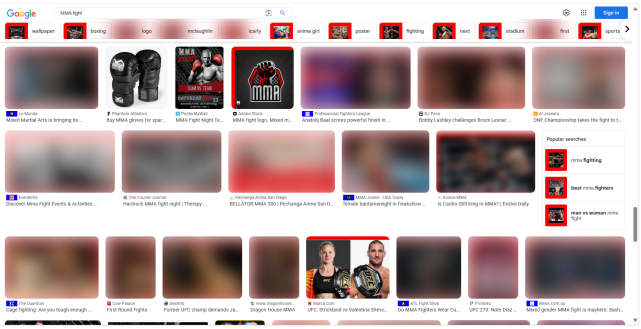

Figure 4. Extension automatically filter “violent" images

The extension necessitates efficient processing times to ensure a seamless user experience. The running time for the model to predict whether an image contains violent content is approximately **2 seconds**. However, due to the concurrent processing capability of the extension, this time is optimized. For the first image, the model takes 2 seconds, but subsequent images benefit from this concurrency. For instance, the second image is processed at 2.4 seconds, the third at 2.8 seconds, and so on. This overlapping processing mechanism ensures that while the first image sets the initial processing time, each subsequent image only adds an incremental 0.4 seconds to the total processing time, thus maintaining efficient and rapid content filtering.

In summary, the extension integrates efficient client-side filtering to block violent images, creating a safer browsing experience. Through minimal filtering, the extension creates a safer online space for underage users, while upholding unconstrained availability of content across the open internet.

VII. CONCLUSION

Through the research and development process for building a new tool for protecting children, we have developed a multi-task learning model capable of detecting and filtering images with violent characteristics regarding the appearance of humans in the images. We first collect sufficient data in many realistic cases to extend the public datasets. The proposed method improves the performance of models regarding diverse cases compared to the single-task model. Thereafter, we successfully applied the model with the binary rule for the web extension to provide flexible and safe management capabilities for children.

Although we have achieved promising results, we recognize that there are still potential research directions that could be further explored such as continuing to increase

realistic training data. We advocate the necessity of combining the detection of sensitive images into our model to provide safer online environments for children.

ACKNOWLEDGEMENT

This work is conducted under the support of Vietnam - Korea University of Information and Communication.

REFERENCES

[1] Unicef-annual-report-2021, https://www.unicef.org/reports/unicef-annual-report-2021. Last accessed 21 Feb. 2024

[2] Van Bruwaene, D., Huang, Q. & Inkpen, D. A multi-platform dataset for detecting cyberbullying in social media. Lang Resources & Evaluation 54, 851–874 (2020).

[3] Van Bruwaene, David, Qianjia Huang, and Diana Inkpen. “A multi-platform dataset for detecting cyberbullying in social media." Language Resources and Evaluation 54 (2020): 851-874.

[4] Ramzan, Muhammad, Adnan Abid, Hikmat Ullah Khan, Shahid Mahmood Awan, Amina Ismail, Muzamil Ahmed, Mahwish Ilyas, and Ahsan Mahmood. “A review on state-of-the-art violence detection techniques." IEEE Access 7 (2019): 107560-107575.

[5] Mumtaz, N., Ejaz, N., Habib, S., Mohsin, S. M., Tiwari, P., Band, S. S., & Kumar, N. (2023). “An overview of violence detection techniques: current challenges and future directions." Artificial intelligence review, 56(5), 4641-4666.

[6] Netnanny, https://www.netnanny.com/. Last accessed 21 Feb. 2024

[7] Cybersitter, https://www.cybersitter.com/. Last accessed 21 Feb. 2024

[8] Maxprotect, https://www.maxprotect.com/. Last accessed 21 Feb. 2024

[9] M. Hammami, Y. Chahir and L. Chen, “WebGuard: Web based adult content detection and filtering system," In Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003), Halifax, NS, Canada, 2003, pp. 574-578.

[10] Hu, Weiming, Haiqiang Zuo, Ou Wu, Yunfei Chen, Zhongfei Zhang, and David Suter. “Recognition of adult images, videos, and web page bags." ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 7, no. 1 (2011): 1-24.

[11] Sharma, Preeti, Manoj Kumar, and Hitesh Sharma. “Comprehensive analyses of image forgery detection methods from traditional to deep learning approaches: an evaluation." Multimedia Tools and Applications 82, no. 12 (2023): 18117-18150.

[12] Soliman, Mohamed Mostafa, Mohamed Hussein Kamal, Mina Abd El-Massih Nashed, Youssef Mohamed Mostafa, Bassel Safwat Chawky, and Dina Khattab. “Violence recognition from videos using deep learning techniques." In 2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS), pp. 80-85. IEEE, 2019.

[13] Human Action Recognition (HAR) Dataset. https://www.kaggle.com/datasets/meetnagadia/human-action-recognition-har-dataset. Last accessed 21 Feb. 2024

[14] Caruana, Rich. “Multitask learning." Machine learning 28 (1997): 41-75.

[15] Sandler, Mark, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. “Mobilenetv2: Inverted residuals and linear bottlenecks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4510-4520. 2018.

[16] Google Chrome extension API, https://developer.chrome.com/docs/extensions. Last accessed 21 Feb. 2024.

Le Kim Hoang Trung is pursuing the B.Eng. degree in Information Technology from the University of Danang, Vietnam - Korea University of Information and Communication Technology. His research interests include software development, machine learning and deep learning.
Email: trunglkh.21it@vku.udn.vn

Nguyen Van Thanh Vinh is pursuing the B.Eng. degree in Information Technology from the University of Danang, Vietnam - Korea University of Information and Communication Technology. His research interests include software development, machine learning and deep learning.
Email: vinhnvt.21it@vku.udn.vn

Phan Le Viet Hung is pursuing the B.Eng. degree in Information Technology from the University of Danang, Vietnam - Korea University of Information and Communication Technology. His research interests include software development, machine learning and deep learning.
Email: hungplv.21ad@vku.udn.vn

Nguyen Huu Nhat Minh (M'20) received Ph.D. degree in Computer Science and Engineering from Kyung Hee University, South Korea, in 2020. He continued Post-Doc with Federated Learning and Democratized Learning at Intelligent Networking lab, Kyung Hee University, South Korea. He is Deputy Head of Department of Science, Technology, and International Cooperation, and In charge of Research Program at Digital Science and Technology Institute, The University of Danang – Vietnam - Korea University of Information and Communication Technology, Vietnam. He received the best KHU Ph.D. thesis award in engineering in 2020. He had publications in premier ACM/IEEE journals and conferences. His research interests include wireless communications, federated learning, NLP, and computer vision.
Email: nhnminh@vku.udn.vn