JTE HCMUTE
ISSN: 2615-9740

JOURNAL OF TECHNICAL EDUCATION SCIENCE
Ho Chi Minh City University of Technology and Education
Website: https://jte.edu.vn
Email: jte@hcmute.edu.vn

Transfer Learning for Abnormal Object Detection

Dung Nguyen
University of Sciences, Hue University, Hue City, Vietnam
Email: nguyendung@hueuni.edu.vn

ARTICLE INFO

Received: 20/01/2024
Revised: 23/02/2024
Accepted: 23/02/2024
Published: 28/02/2024

KEYWORDS

Machine learning;
Transfer learning;
Finetuning;
Feature Extraction;
Object Detection.

ABSTRACT

In today's world, smart surveillance plays an important role in protecting security and creating a safe living environment. For abnormal objects in the smart surveillance system, this is an important issue, requiring attention and timely response from managers and supervisors. To address this issue, the paper uses transfer learning techniques on modern object detection models to detect abnormal objects such as guns, knives, etc. in public places. We experimented with the transfer learning method on the DETR model with a small dataset, and the model results showed a fairly fast convergence speed. Through this method, we hope to help reduce the burden of public security monitoring and warning work for managers, while technicians can use transfer learning techniques that are deployed in practice.

Doi: https://doi.org/10.54644/jte.2024.1526

1. Introduction

Although machine learning- based traditional methods or even deep learning methods have reached great success and are successfully applied in a lot of real-world application systems, still has certain limitations in the training process, such as: computing infrastructure, data size, labeled data, and training time. In many cases, collecting enough training data that is usually expensive, time-consuming, even impractical. The initial solution proposed was to use a semi-supervised learning method. This approach can solve part of this problem by reducing the requirement for labeled dataset. Semi supervised learning methods only need a small amount of labeled data, and it uses a large amount of unlabeled data to improve learning accuracy. In many situations, unlabeled data samples are also difficult problems to collect, which often makes the resulting machine learning models unsatisfactory. Therefore, transfer-based learning methods were proposed. Transfer learning is a learning method that transfers knowledge between application domains, from one model to another, from one task to another, and is a high potential machine learning method to deal the above problems [1].

Fig 1. *Traditional learning vs transfer-based learning approaches*

The transfer-based learning is a machine learning approach, which supports to train a model from one task and then applying that model to a new task, or new data domain [1], [2]. This is typically done

using a pre-trained model on a large, high-coverage dataset. The pre-trained model can then be fine-tuned on a smaller dataset for a new task [3]. Transfer learning can be understood as using a model's prior knowledge and experience to learn a new task, see illustration in fig. 1. For example, a model that has been trained to recognize objects can be used to learn how to recognize license plates. In this case, the model's knowledge and experience on shape and feature recognition can be used to learn license plate recognition.

The transfer-based learning approach can be used to improve a model's performance on a new problem, the amount of data available for the new task is limited, especially. It can also be used to reduce the amount of time and computational resources required to train a model. Here are some examples of transfer learning being used in the real world:

(1) Image classification [4]-[7]: Transfer learning is often used to train image classification models. For example, a model trained using ImageNet is utilized to classify images of new objects, such as medical images or satellite images.
(2) Natural Language Processing [7]: Transfer learning is also used to train NLP models. For example, a model that is trained on a dataset of the large corpus of text is used to perform tasks such as sentiment analysis or machine translation.
(3) Recommender system [8], [9]: Transfer learning is utilized to improve the efficiency of the recommender system. For example, a model that has been trained to recommend movies to users can be used to recommend books to users.

The transfer-based learning method is a powerful approach that is utilized to improve the performance of machine learning models. It is a versatile technique that can be applied to many types of tasks. As the amount of data available to train machine learning models continues to increase, transfer learning may become an even more important tool for machine learning professionals. Transfer-based learning can provide many benefits, including:

(1) Improved performance: Transfer - based learning method supports improving the model performance for new tasks, especially the new task with small data.
(2) Reduce time and cost: Transfer learning can help reduce the time and cost needed to train a model.
(3) Expand the scope of application: Transfer learning support to expand the scope of application of machine learning models.

Some popular Transfer Learning techniques used in the field of machine learning:

(1) Fine-Tuning: The pretrained model on a large dataset and then fine-tuned on a new dataset often used when the target dataset is not large enough and is like the source dataset.
(2) Feature Extraction: Use the layers of the previously trained model to extract features from the data and then feed it into a new model to train new classification layers. Effective when the target dataset is small, and the classification task is similar to the source dataset.
(3) Pre-training: The model is pre-trained on a large and multi-task dataset, usually on a common dataset or a large dataset from the internet. Often used to learn general features and transfer knowledge from complex tasks to specific tasks.
(4) Domain-specific Pre-training: The model is pre-trained on a large and multi-task dataset but focuses on a specific domain or context. Useful when the target dataset belongs to a specific domain and wants to transfer knowledge from that domain.
(5) Multi-task Learning: The model is trained on multiple tasks at the same time, sharing some classes or feature extraction. Can help improve model performance across different tasks and transfer knowledge from one task to another.
(6) Self-supervised Learning: The model automatically generates fake data from available data and then uses it to train the model. The model is then transferred to the target dataset. Effective when there is not enough labeled data and want to take advantage of unlabeled data.

These techniques can be combined or adapted depending on the specifics of the data and the task we are working with.

2. Related Works

In this article, we introduce some popular object detection models. Object detection models can be divided into one-stage models and two-stage models. Two-stage models are models that try to search for an arbitrary number of proposed objects in an image in the first stage. The task of classification and localization is the second stage. The models have two separate steps, they often take more complexity to generate object container proposals with complex architectures. The one-stage detection model classifies and locates objects in a single pass using a dense sampling method. It has a simpler design and has higher real-time performance than two-stage detection models.

R-CNN model family. The region-based convolutional neural network model family originated with the introduction of the R-CNN model, a groundbreaking paper by Girshick and colleagues in 2012 [10]. This marked the beginning of using CNN to significantly enhance object detection performance. R-CNN employs a selective search method to suggest regions containing Region of Interest (RoI) objects, utilizing a CNN network to extract features. The image undergoes a region proposal module, generating around 2000 region objects using the Selective Search algorithm [11] to identify parts of the image with a higher likelihood of containing the object. These RoIs are then processed through a CNN network, such as the AlexNet [12] as the backbone architecture in Girshick et al.'s implementation, generating a 4096-dimensional feature vector for each RoI. A trained class-specific support vector machine (SVM) is then employed for classification, determining whether it belongs to the corresponding class. The algorithm estimates bounding boxes using a trained bounding box regressing, estimating the center coordinates, width, and height.

While R-CNN revolutionized object detection, it suffered from slow processing and high computational costs. To address these drawbacks, the Fast R-CNN [13] model was introduced in 2015 by Girshick and colleagues as an improvement in speed, achieving 146 times faster processing than R-CNN. Although Fast R-CNN approached real-time object detection, its RoI region proposal generation remained relatively slow. Instead of the Selective Search algorithm, the authors introduced the Region Proposal Network (RPN) [14] to find RoIs, replacing the earlier algorithm.

Faster R-CNN was proposed [15], further refined the RPN concept. This algorithm employs multiple bounding boxes with different aspect ratios, regressing them to determine the object's location. The input image first passes through CNN to obtain feature maps, which are then forwarded to the RPN. The RPN generates region or bounding box proposals along with their classification. The selected RoI region proposals are mapped back to the feature maps, which were obtained from the previous CNN layer. As a result, Faster R-CNN significantly is better Fast R-CNN in terms of speed, representing a notable advancement within the R-CNN model family.

Transformer model family. In recent years, the Transformer paradigm has profoundly influenced the entire field of deep learning, especially the field of computer vision. The new approach based on transformer model [16] eliminates the traditional convolution operator and instead only calculates based on the self-attention mechanism to overcome the limitations of CNN. In 2020, N. Carion et al proposed DETR [17], which proposed an end-to-end detection network with Transformers. This model uses a CNN network to extract image features. Here the author uses the ResNet [18] network, as a result we obtain a feature map of the input image. This feature map is added to the position encoding vector, to determine the order of features. The result of this addition is transferred to the Transformer's Encoder network for encoding. The results of the Encoder network (after doing it 6 times) are fed into the Decoder network for decoding. As the output of the Decoder network, we obtain a feature map, which we use to classify and determine bounding-boxes through the Full connected, Soft-max and regression layers.

YOLO model family. Starting with its inaugural version, YOLOv1 (You Only Look Once), represents a pioneering approach to object detection utilizing deep neural networks for recognizing and localizing objects within images and videos. Introduced in 2016 by Joseph Redmon and Santosh Divvala [19], YOLOv1 revolutionized computer vision's object detection capabilities by dividing the input image into an S × S grid. Each grid cell is tasked with detecting an object if its center falls within that cell. However, a drawback of YOLOv1 is its limitation in detecting small, overlapping objects, and it can predict a maximum of S x S objects in the image.

Building upon the foundation of YOLOv1, YOLOv2 [20] (also known as YOLO9000) was introduced in 2016 as an enhancement over the original YOLO algorithm. YOLOv2 aims to be faster and more accurate, featuring the use of anchor boxes, which predefined bounding boxes with many scales and aspect ratios. This improvement enables YOLOv2 to detect a wider variety of objects.

YOLOv3, presented in 2018 [21], further elevates the YOLO object detection algorithm's accuracy and speed. Notable enhancements include the adoption of the Darknet-53 CNN architecture, a variant of ResNet specifically tailored for object detection tasks with 53 convolutional layers. YOLOv3 introduces anchor boxes with varied scales and aspect ratios, adapting to the size and shape of detected objects. The "feature pyramid network" (FPN) is also introduced, facilitating object detection at multiple scales and improving performance on small objects.

Moving forward to YOLOv7, unveiled in a 2022 paper [22], significant improvements are implemented. YOLOv7 leverages anchor boxes—predefined boxes with different aspect ratios—to detect objects of various shapes. With 9 anchor boxes, YOLOv7 broadens its capability to detect a diverse range of object shapes and sizes, minimizing false positives. Additionally, YOLOv7 boasts higher resolution compared to its predecessors.

While YOLOv8, developed by the Ultralytics team, is a recent addition to the YOLO model family, detailed papers about its architecture are yet to be published. Nonetheless, the computer vision community has access to and has been testing the model using the source code shared on the team's GitHub page. YOLOv8 is an advanced computer vision model with built-in support for object detection, classification, and segmentation tasks, extending the capabilities of its predecessors.

3. Solution for Abnormal Object Recognition

There are two main types of transfer learning [23]:

(1) Intuitive transfer learning [24]: This is the most common type of transfer learning, and it involves using a pre-trained model to learn a new task which is related to the original task. The task is demonstrated in Fig. 2. For example, a model trained to recognize cats and dogs can be used to learn to recognize new animals, such as horses and cows.

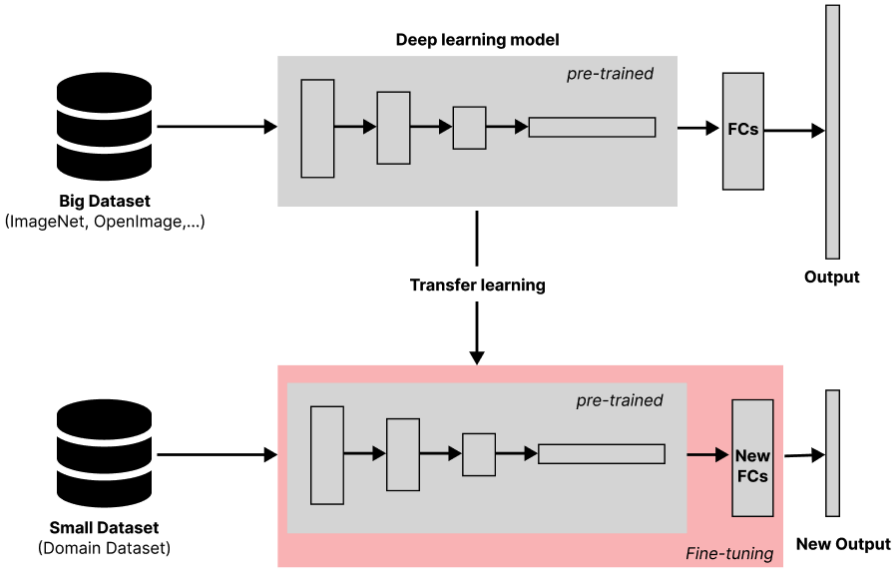


Fig 2. *Inductive transfer learning*

(2) Deductive transfer learning: This type of transfer learning involves using a pre-trained model to learn a new task unrelated to the original task as shown in Fig. 3. For example, a model that has been trained to translate French to English can be used to translate Spanish to English.

JTE HCMUTE
ISSN: 2615-9740
JOURNAL OF TECHNICAL EDUCATION SCIENCE
Ho Chi Minh City University of Technology and Education
Website: https://jte.edu.vn
Email: jte@hcmute.edu.vn

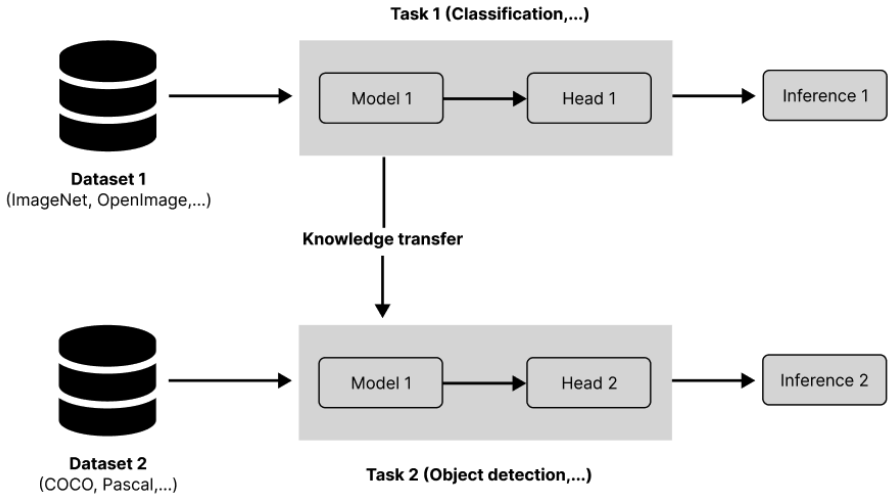


Fig 3. *Deductive transfer learning*

The method we employed in this paper involves conducting experiments using the DETR architecture that has been pre-trained on the large COCO dataset. In this approach, we performed Finetuning on the output of the DETR model using the COCO dataset to adapt it to the output on the dataset we experimented with. This method can be illustrated as shown in Figure 4.

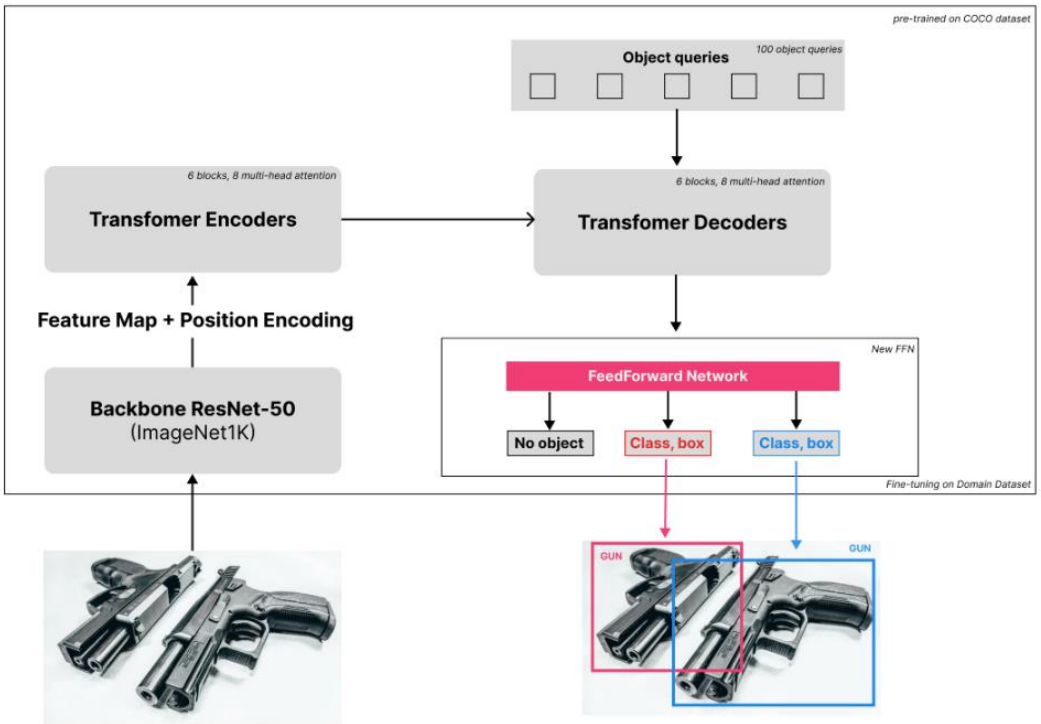


Fig 4. The transfer – based learning approach for object detection systems

4. Experiments and Results

We train the DETR model on the WEAPON dataset using a model previously trained with the large dataset COCO 2017. The details are described below:

Dataset: We conduct experiments on the WEAPON dataset [25] as described in table 1. This data set includes 22212 images, divided into 3 subsets, of which the training set includes 19434 images.

Table 1. *WEAPON dataset*

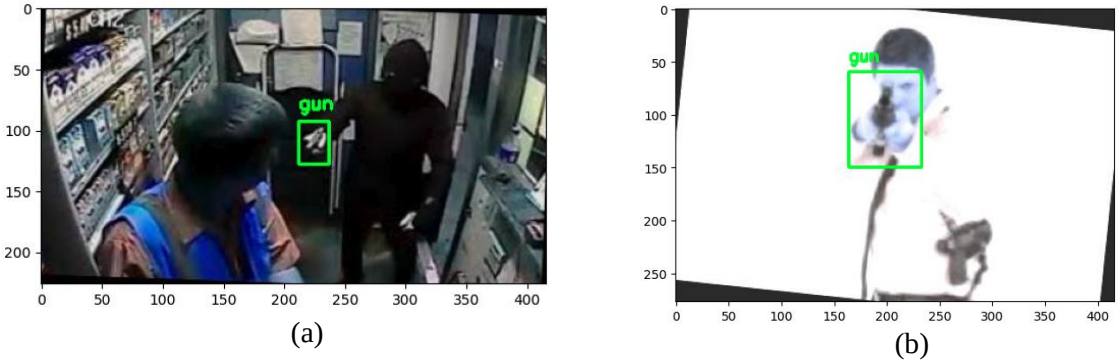
Data	Images	Ratio	Augmentations
Total	22212	100%	Flip: Horizontal
Train	19434	88%	Crop: 0% Minimum Zoom, 35% Maximum Zoom
Val	1851	8%	Rotation: Between -8° and +8°
Test	927	4%	Brightness: Between -25% and +25%

Implement Details. To train the model, we keep the parameter set of the original DETR model as shown in table 2. We use the parameter set of the already trained and published model as facebook/detr-resnet-50.

Table 2. *Training parameter set*

Parameter	Value
Encode, Decoder layers	6
Backbone	Resnet-50
Pre-train BackBone	ImageNet1K
Optimizer	AdamW
Batch_size	2
Learning rate	1e-4
Epoch number	50

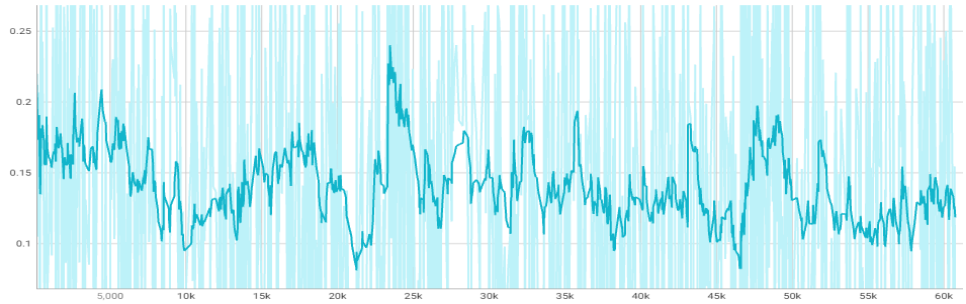
Some images in the dataset: Some illustrative images in the dataset are shown in Fig. 6.



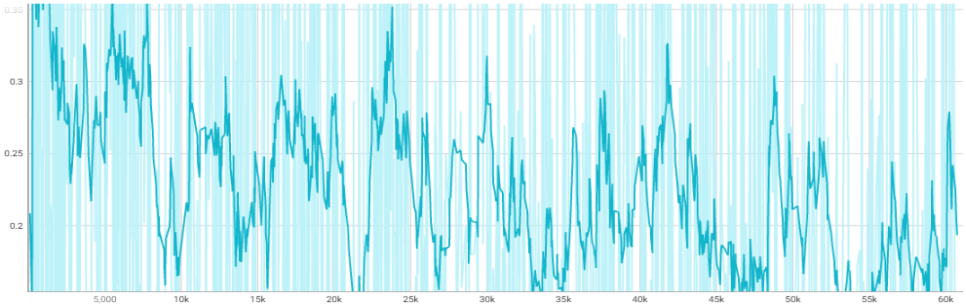
(a) (b)

Fig 6. *Some images in the dataset*

Results on the train dataset. Experimental results on the training data set are as described in Fig. 7.



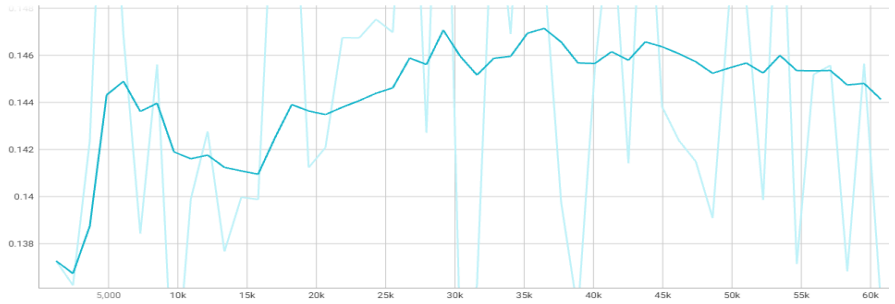
(a) Train loss CE (Cross-Entropy)



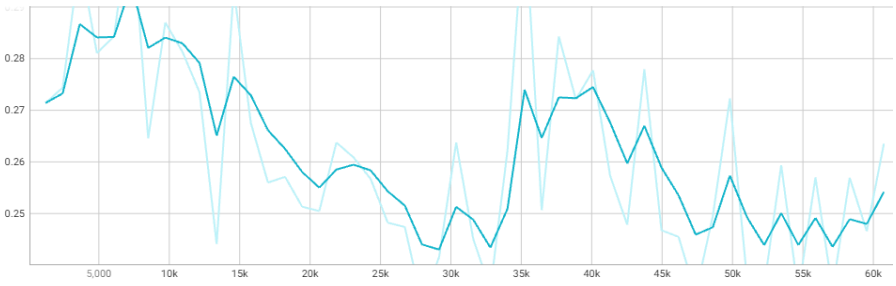
(b) Train loss GIOU

Fig 7. *Experimental results on Train Dataset*

Results on the valid dataset. Experimental results on the validation dataset are described in Fig. 8.



(a) Valid loss CE (Cross-Entropy)



(b) Valid loss GIOU

Fig 8. *Experimental results on Valid Dataset*

5. Conclusion

Transfer Learning is a powerful learning technique that helps solve many major problems in the field of machine learning and computer vision such as lack of data, training time and training infrastructure. The article summarizes the main points and tests this technique on a model for the task of object detection to evaluate the effectiveness and accuracy of the models when applying this technique. The

results show that applying transfer learning techniques helps the model learn faster and with better accuracy than training the model using conventional techniques. This helps scientists open up the right direction in using existing models for use in new data domains and new tasks. At present, we have only conducted experiments with this method on a dataset using one object detection architecture, so there are not many results to compare with the outcomes of transfer learning on different architectures. In the future, we will experiment with more models and a broader range of datasets, enabling us to compare the advantages and disadvantages of this transfer learning approach across various architectures.

Conflict of Interest

The authors declare no conflict of interest.

REFERENCES

[1] F. Zhuang *et al.*, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43-76, 2020.
[2] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, pp. 1-40, 2016.
[3] A. Pasini, "Artificial neural networks for small dataset analysis," *Journal of thoracic disease*, vol. 7, no. 5, p. 953, 2015.
[4] M. Bansal, M. Kumar, M. Sachdeva, and A. Mittal, "Transfer learning for image classification using VGG19: Caltech-101 image data set," *Journal of ambient intelligence and humanized computing*, pp. 1-12, 2021.
[5] M. Shaha and M. Pawar, "Transfer learning for image classification," in *2018 second international conference on electronics, communication and aerospace technology (ICECA)*, 2018: IEEE, pp. 656-660.
[6] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: a literature review," *BMC medical imaging*, vol. 22, no. 1, p. 69, 2022.
[7] N. Agarwal, A. Sondhi, K. Chopra, and G. Singh, "Transfer learning: Survey and classification," *Smart Innovations in Communication and Computational Sciences: Proceedings of ICSICCS 2020*, pp. 145-155, 2021.
[8] L. Zhao, S. Pan, E. Xiang, E. Zhong, Z. Lu, and Q. Yang, "Active transfer learning for cross-system recommendation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2013, vol. 27, no. 1, pp. 1205-1211.
[9] Z. Lin, D. Liu, W. Pan, and Z. Ming, "Transfer learning in collaborative recommendation for bias reduction," in *Proceedings of the 15th ACM Conference on Recommender Systems*, 2021, pp. 736-740.
[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580-587.
[11] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, pp. 154-171, 2013.
[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.
[13] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440-1448.
[14] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904-1916, 2015.
[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.
[16] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
[17] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*, 2020: Springer, pp. 213-229.
[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.
[19] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779-788.
[20] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263-7271.
[21] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.
[22] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7464-7475.
[23] A. Hosna, E. Merry, J. Gyalmo, Z. Alom, Z. Aung, and M. A. Azim, "Transfer learning: a friendly introduction," *Journal of Big Data*, vol. 9, no. 1, p. 102, 2022.
[24] H. D. Nguyen and C. Sakama, "Feature learning by least generalization," in *International Conference on Inductive Logic Programming*, 2021: Springer, pp. 193-202.
[25] *Gun Detectiongun Dataset*, Roboflow [Online]. Available: https://universe.roboflow.com/ruclan99999-mail-ru/gun-detectiongun

Nguyen Dung was born on June 13, 1988 in Thua Thien Hue. He graduated with a bachelor's degree in information technology from the College of Sciences, Hue University in 2010. In 2013, he graduated with a master's degree in computer science from the College of Sciences, Hue University. Currently he works at the University of Sciences, Hue University. Research fields: Software technology, artificial intelligence, machine learning, deep learning, databases

Email: nguyendung@hueuni.edu.vn. ORCID: https://orcid.org/0009-0000-4510-7504