An Approach for Object Recognition in Videos for Vocabulary Extraction

Anh Bao Nguyen Le[1], Chi Bao Nguyen[1], Quoc Cuong Dang[1], Be Hai Danh[1], Huynh Nhu Le[1], Huong Hoang Luong[2], and Hai Thanh Nguyen[1(✉)]

[1] College of Information and Communication Technology, Can Tho University, Can Tho, Vietnam
nthai.cit@ctu.edu.vn
[2] FPT University, Can Tho, Vietnam

Abstract. English is the most common language globally, and it is increasingly important. English has been compiled in most online documents, information, and contents. However, with a considerable vocabulary, learning English is difficult for many people to remember. Therefore, many modern technologies have been proposed to support English learning, such as English learning technology through word-matching games to help children become excited and easily approach English from an early age. In addition, translation tools can help users look up vocabularies, antonyms, synonyms, and examples. This study presents a method to support learning English via object detection in videos, images, or even live-stream videos in real-time using deep learning architectures such as You Look Only Once (YOLO) - one of the finest families of object detection models with state-of-the-art performances. The method to obtain an mAP is 55.6 with 17GFlops. The results are vocabulary, meaning, and making sentences with that. Our method has good accuracy in data of 2786 images belonging to 59 classes.

Keywords: Vocabulary · learning English · object detection

1 Introduction

English has such an important role that society's requirements for proficiency in the English language are high. However, looking deeply into that problem, firstly, on the student's side, we may understand what the question is asking when participating in big and small exams. Secondly, for those working in professions needing to communicate fluently in English, the reality of their foreign language ability is still too, and the need to improve their vocabulary is limited, hindering their ability to meet the community's needs. Applicable in current and future work communication. Third, parents always want their children to start learning English from 3 to 4 years old, but they have a headache because they need to learn how to teach their children the basics. Therefore, it is necessary to have a method of learning English from basic to advanced.

P. Cong Vinh and H. Mahfooz Ul Haque (Eds.): ICTCC 2023, LNICST 586, pp. 36–51, 2024.
https://doi.org/10.1007/978-3-031-59462-5_3

As the importance of English vocabulary gradually asserts its position, the search for long-term and effective learning methods is promoted. Therefore, learning English vocabulary through pictures has become popular and strongly developed. That is also the method our team chose to research to help people improve their English vocabulary. Our brains tend to remember images and words more. Memorizing words that appear simple will make learning more exciting and not dull. At the same time, it helps to stimulate the brain, which will help us remember longer. As students majoring in information technology at Can Tho University, we are dedicated to successfully researching an Android application to learn English vocabulary through images and videos, applying AI(Artificial Intelligence) science and API(Application Programming Interface) to research, and learning additional guidance.

Because mobile apps are becoming increasingly important in our lives, this research will make vocabulary learning more enjoyable, effective, and memorable. Furthermore, users can learn for free anytime and from any location. They require a smartphone. As a result, it will contribute significantly to the advancement of education.

Furthermore, the primary research clarifies the role of information technology application in human life in general and in English learning in this article. From there, technology will serve as a bridge for Vietnamese technology to advance further by assisting people in their knowledge development. Artificial intelligence, or AI, is a branch of computer science that refers to intelligence humans have programmed to assist computers in automating intelligent human-like behaviors. In particular, artificial intelligence assists computers in absorbing human intelligence, such as image recognition, voice recognition, inference to solve problems, etc. Many people are becoming increasingly interested in the research and application of AI technology. Moreover, real-world application In the case of AI(Artificial Intelligence), image recognition, for example, uses Roboflow to label an image, select it, and label it. For example, if we train a cat image with the label "cat," the cat image will display the English word "cat" after encountering a specific image related to the voice used-a programming interface API(Application Programming Interface) for reading text, specifically vocabulary. The API is intended to enable text-to-speech functionality. An artificial intelligence-based English learning application or software was designed to assist English learners in learning vocabulary and pronunciation through intellectual activities. Since the introduction of English learning applications, the trend in learner acceptance and effective learning methods has shifted. The rapid advancement of technology allows popular learning applications to be downloaded to phones, laptops, and mobile devices on all platforms, including Android, iOS, Windows, etc.

As the importance of English vocabulary gradually increases, the search for long-term and effective learning methods is promoted. Therefore, learning English vocabulary through pictures has become popular and strongly developed. That is also the method our team chose to research to help people improve their English vocabulary. Our brains tend to remember images and words more. Memorizing words that appear simple will make learning more exciting and

not dull. At the same time, it helps to stimulate the brain, which will help us remember longer. As students majoring in information technology at Can Tho University, we are dedicated to successfully researching an Android application to learn English vocabulary through images and videos, applying AI(Artificial Intelligence) science and API(Application Programming Interface) to research and learn additional guidance. Because mobile apps are becoming increasingly important, this research will make vocabulary learning more enjoyable, effective, and memorable. Furthermore, users can learn for free anytime and from any location. They require a smartphone. As a result, it can contribute significantly to the advancement of education. Furthermore, the primary research clarifies the role of information technology application in human life in general and in English learning in this article. From there, technology will serve as a bridge for Vietnamese technology to advance further by assisting people in their knowledge development. Artificial intelligence, or AI, is a branch of computer science that refers to intelligence humans have programmed to assist computers in automating intelligent human-like behaviors. In particular, artificial intelligence assists computers in absorbing human intelligence, such as image recognition, voice recognition, inference to solve problems, etc. Many people are becoming increasingly interested in the research and application of AI technology. Moreover, real-world application In the case of AI(Artificial Intelligence), image recognition, for example, uses Roboflow to label an image, select it, and label it. For example, if you train a cat image with the label "cat," the cat image will display the English word "cat" after encountering a specific image related to the voice used-a programming interface API(Application Programming Interface) for reading text, specifically vocabulary. The API is intended to enable text-to-speech functionality. An artificial intelligence-based English learning application or software was designed to assist English learners in learning vocabulary and pronunciation through intellectual activities. Since the introduction of English learning applications, the trend in learner acceptance and effective learning methods has shifted. The rapid advancement of technology allows popular learning applications to be downloaded to phones, laptops, and mobile devices on all platforms, including Android, iOS, Windows, etc. Object detection has become a massive machine-learning field in recent years. Video object detection is no exception; it is still new but increasingly important in our lives. So what is video object detection, and how important is it? Information Technology researchers and developers have decided to create Video Object Detection applications that allow machines to analyze images and detect objects. Video object detection helps reduce operational human resources, has high accuracy, continuous operation output, and is easy to monitor and operate, so it has attracted many technology industries to apply this method: number plate recognition, face recognition detection, object tracking, cars - self-driving aircraft, robotics, etc. Many prospective studies show the importance of VOD in the future being implemented in many areas such as sports (player identification and analysis of broadcast video about football [1]); medicine (detection of microscopic objects through microscopic video and analysis of sperm quality [2,3]); security (automatic gun detection [4]); space science

(detecting and tracking moving objects in satellite video [5]), etc. We researched and developed a VOD in the field of education based on the benefits and importance mentioned above. It is image recognition via video conversion into English vocabulary (English vocabulary learning software). Only in the field of education, in this day and age of technology, with the explosion of the internet, is it common for students to own a phone, laptop, or another electronic device. The software is simple to install on the learner's learning device. It is critical to raise the level of education in the country to create closeness and comfort for any age group, including parents who wish to teach their children at home. It can also help people perceive technology in their daily lives more quickly. As a result, technology engineers will be inspired and driven to develop innovative and practical software.

The rest of this paper is organized as follows. In Sect. 2, the related work is introduced. In Sect. 3, the proposed fault diagnosis method is presented in detail. In Sect. 4, detailed experiments and comparisons are carried out. The summary of this paper is presented in Sect. 5.

2 Related Work

Video object detection (VOD) has gained popularity in recent years. Many studies have been conducted. Video object detection is actively studied to push detection speed and accuracy limits. The authors in [6] proposed that the method extracts a set of convolutional feature maps over the whole input image via a fully convolutional backbone network and performs region classification and bounding box regression over either sparse object proposals. The method achieved the mapping score of 78.6% at runtime of 13.0/8.6 fps on Titan X/K40, better when compared to ImageNet VID Challenge 2017 with an mAP score of 76.8% at runtime of 15.4 fps on Titan X, shows a progressive result, towards high-performance video object detection.

The authors in [7] suggested a cuboid proposal network and tubeless linking algorithm to improve the performance of detecting moving objects in videos. Experiments on the ImageNet VID dataset show that their method outperforms the static image detector and the previous state-of-the-art. In particular, their method improves results by 8.8 percent over the static image detector for fast-moving objects. Another work in [8] introduced an object query propagation (QueryProp) framework for high-performance video object detection.

They evaluated their model on ImageNet VID, which consists of 3862 training videos and 555 validation videos from 30 object categories. QueryProp propagates sparse object queries across video frames to achieve online video object detection, and no additional modules or post-processing are required. The processing speed of QueryProp can achieve 45.6 FPS while maintaining an accuracy of over 80 mAP. This novel solution enables a new framework to achieve the best performance among all online video object detection approaches and strikes a decent accuracy/speed trade-off.

The work in [9] designed the Hierarchical Video Relation Network (HVR-Net), which uses inter-video and intra-video proposal relations to improve object

feature quality. They mainly evaluate their HVR-Net HVR-Net on the large-scale ImageNet VID dataset. It comprises 3862 training videos (1,122,397 frames) and 555 validation videos (176,126 frames), with bounding box annotations across 30 object categories. HVR-Net was influential and essential for video object detection.

In addition, industries such as autonomous driving, surveillance systems, drones, and robotics are increasingly driven by tremendous success in the VOD sector. For example, autonomous driving leverage video recognition, whose market is predicted to leap to $77 billion (25% of the whole automotive market) by 2035 [10], has attracted the attention of giants including Tesla[1] and Waymo[2]; surveillance systems are applied in many aspects: intelligent transportation, intelligent oil field production, and management optimization, water conservation monitoring with Advantech WebAccess automation,... [11] Along with that development, deep neural network (DNN) techniques are vulnerable to adversarial attacks. For the above reason, Themis is a software/hardware system to defend against adversarial patches for real-time robust video object detection that was recently launched [12]. Themis efficiently and accurately recovers the DNN systems from adversarial attacks with the algorithmic framework and architectural support. The results show that the proposed methodology can recover the VOD system's negative effect in real time with negligible hardware overhead.

Based on the positives regarding detection speed and accuracy, many integration studies use VOD in many practical applications. For example, the authors in [13] have researched a method that presented a fully automated pipeline for face detection, tracking using a deep convolutional neural network (CNN). In addition, a fast car detection and tracking algorithm was presented in [14] for traffic monitoring fisheye video mounted on crossroads. They used the ICIP 2020 VIP Cup dataset and adopted YOLOv5 as the object detection base model.

They studied 26 videos for training and five videos for testing, taken from a fisheye camera mounted on a pole near road intersections about 8 m above the road. Each video typically has around 1000 frames captured at 15 frames per second. Their design improves the detection rate by 17.9 pp in the night scenes and 6.2 pp for the day scenes, increasing the inference speed by nearly two times. The authors in [15] also performed a challenging task of object-based video forgery detection. They used the fast and real-time object detector You Only Look Once (YOLO) -Version 2 to automatically detect forged objects within the video with a 0.99 confidence score. A study focuses on object detection from thermal infrared images and videos of UAVs using the YOLO models once deployed in [16]. Object detection has been performed on various remote sensing platforms on spaceborne, aerial, and ground remote sensing images and videos. Results revealed that the highest mean average precision (mAP) of the person and car instances was 88.69%, the fastest detection speed achieved 50 frames per second (FPS), and the smallest model size was observed in YOLOv5-s. Recent

[1] https://electrek.co/2017/04/29/elon-musk-tesla-plan-level-5-full-autonomous-driving/.

[2] https://blog.waymo.com/2019/08/introducing-waymos-suite-of-custom.html.

learning-based video methods (e.g., [17–19]) typically require an extensive collection of well-annotated data for learning a new object class, making it difficult to scale to real-world object classes in high diversity. Therefore, Few-Shot Video Object Detection (FSVOD) was introduced in [20] with three critical contributions: a large-scale video dataset FSVOD-500 comprising 500 classes with class-balanced videos in each category for few-shot learning; a novel Tube Proposal Network (TPN) to generate high-quality video tube proposals to aggregate feature representation for the target video object; a strategically im-proved Temporal Matching Network (TMN+) to match representative query tube features and supports with better dis-criminative ability. Extensive experiments demonstrate that their method produces significantly better detection results on two few-shot VOD datasets, boosting FSVOD research potential in the future.

3 Method

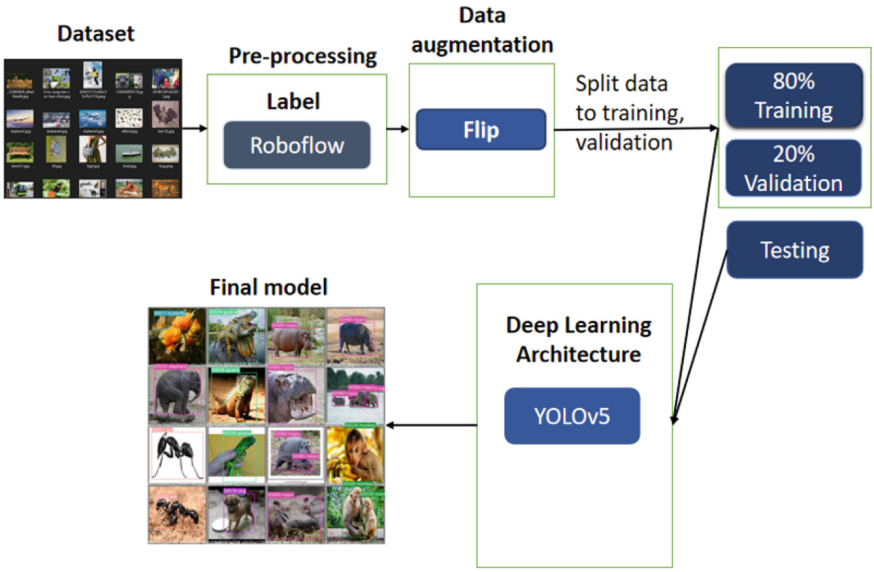

Fig. 1. The workflow for object detection in images

Figure 1 illustrates the process of forming the labeled object. The first image, from left to right, is an image database with more than 2,500 images used to label the correct vocabulary with the meaning of each image. The second figure, Roboflow, is a framework with a Label Assistant feature, where labels can be applied to objects or predictably for faster labeling. Next is using Data Augmentation with one primary method: flip. In splitting the data into 80% training and 20% validation, the testing set is not initially included in the dataset. Next is to pass the found model through machine learning, using two main types of machine learning architectures, YOLOv5, to classify images by taking an input image,

processing and analyzing it, and categorizing it under specific categories (Example: Butterfly, Elephant...), and finally, provide the classification and detection results.

3.1 Data Pre-processing

The image processing system's input is images taken from many different sources, including data sets on the internet and phone cameras. Image files are usually of good quality because they have gone through careful selection steps to save time and increase the quality of the process.

Fig. 2. Some examples of Labeling the pictures using Roboflow

We have collected 2786 images of 59 classes in Can Tho City, Vietnam, and some images from the internet. The number of images of each layer depends on that class's variety of shapes, colors, and shooting angles. These classes are

popularly-appeared objects found in Can Tho City, Vietnam. The resolution of the image is primarily good due to careful selection.

Labeling images are carefully labeled due to regular checks to improve processing quality. Image labeling is being done on the website "Roboflow"[3] as illustrated in Fig. 2. Roboflow creates software products as a service that helps users manage image files, annotations, labels, pre-processing, data augmentations, file formats, and model training with one click. Make computer vision tasks easier.

3.2 Data Augmentation

Data augmentation is a technique used in deep learning to improve the data quality used for training, And this time, we use one data augmentation method: Flip. Flip data augmentation is called flip data augmentation by inverting entire rows and columns of image pixels horizontally. For example, in Fig. 3, the horizontal flip is performed on the input image (left side) and returns the image (right side) after flipping the entire pixel of an image.

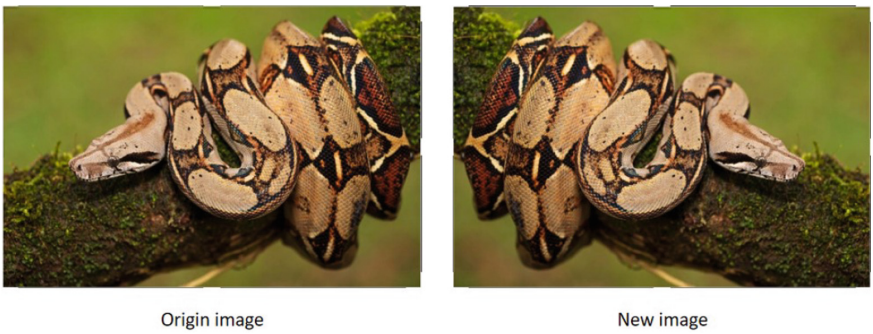

Fig. 3. Flip data augmentation

This way, the data will be increased without adding completely new images. Using the existing image data, flip them, thereby increasing the model's accuracy. Yolov5 [21] is a product within the Yolo architecture series. This model boasts high detection accuracy and fast recognition speed, with the fastest detection speed reaching up to 140 frames per second.

3.3 Object Detection and Classification

Image classification[4] is a complex process, the accuracy of which is based on the dataset's characteristics, the complexity of the problem under analysis, and the

[3] https://roboflow.com.
[4] https://www.sciencedirect.com/topics/earth-and-planetary-sciences/image-classification.

appropriate of the classification algorithm. However, This process is unsuitable for this project because it cannot scan multiple objects in an image. Object detection Image classification is a complex process, the accuracy of which is based on the dataset's characteristics, the complexity of the problem under analysis, and the appropriateness of the classification algorithm. However, This process is unsuitable for this project because it cannot scan multiple objects in an image[5].

Object detection[6] is technique for indicating instances of objects in images or videos. Object detection locates each object by bounding box, and it will classify the object inside each bounding box. As such, there may be more than one object in one image. So it will be suitable for this project[7]. There are different ways to perform object detection. Popular deep learning-based approaches using YOLO v5 automatically learn to detect objects within images[8]. YOLOv5[9] is a family of object detection architectures and pre-trained models on the COCO dataset. The YOLO family of models consists of three main architectural blocks i) Backbone, ii) Neck, and iii) Head.

In our study, we use Yolov5 [21], a product within the Yolo architecture series. This model boasts high detection accuracy and fast recognition speed, with the fastest detection speed reaching up to 140 frames per second. Furthermore, the model's post-training file size is significantly smaller, nearly 90% smaller than Yolov4. This makes it well-suited for deployment on embedded devices for real-time object detection. Yolov5 comprises a total of 10 different architectures, including YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5x, YOLOv5n6, YOLOv5s6, YOLOv5m6, YOLOv5l6, and YOLOv5x6.

4 Experimental Results

We collected 2786 images belonging to 59 classes. Each class has about 30 images with classes mentioned in Table 1. The classes are diversely distributed in all fields and topics. It will update more classes in any fields and topics in the future. The data is divided into two train/valid sets with a ratio of 4/1 and used five pictures per class to be a test set to evaluate the model as exhibited in Table 2. In Sect. 4.2, we train two models with and without data augmentation. The origins origin model has five classes: ant, bear, buffalo, bee, and boa. After using data augmentation shows, the ratio between the Training Set, Validation Set, and Testing Set is shown in the Table 3. In Sect. 4.3, we add two new classes, pen and lip stick, to test the model's ability to update. Each class has 50 samples. We will test the model's training process based on loss to see its ability.

[5] https://kikaben.com/object-detection-vs-image-classification/#chapter-1.
[6] https://www.mathworks.com/discovery/object-detection.html.
[7] https://kikaben.com/object-detection-vs-image-classification/chapter-1.
[8] https://www.mathworks.com/discovery/object-detection.html.
[9] https://github.com/ultralytics/yolov5.

Table 1. Classes in datasets

airplane	butterfly	fire hydrant	pack back	squirrel	peacock
ant	cat	fly	parking meter	starling	rabbit
bat	cheetah	frisbee	parrot	stop sign	raven
bear	chicken	giraffe	pig	suitcase	salamander
bee	cockroach	goat	sheep	tennis ball	scorpion
bench	cow	handbag	skis	tie	snake
boa	crocodile	horse	snowboard	traffic light	squid
boat	dog	iguana	soccer ball	train	stork
buffalo	eagle	monkey	sparrow	truck	hippo
bus	elephant	owl	spider	umbrella	

4.1 Environmental Settings and Metrics

Use web roboflow to label objects and Colab's GPU to run the experimental environment and train in 200 epochs (from 0–199).

Table 2. Top 10 classes with the highest numbers of samples.

Elephant = 100	Dog = 100	Horse = 100	Butterfly = 100	Chicken = 100
Pig = 87	Cat = 50	Sheep = 50	Sparrow = 50	Cow = 50

Table 3. Split the data into training, validation, and test dataset in scenario 3.

Dataset	Original samples	Data after Augmentation
Training	187	351
Validation	24	24
Test	46	46

The performance of approaches is assessed using the F1-score as revealed in Eq. 1, where TP denotes True Positive, FP denotes False Positive, and FN denotes False Negative.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{1}$$

4.2 Efficiency of Data Augmentation

After training two different models with the same class, one is the original model, and the model applied one way of data augmentation: flip. We compare the results of these two models below. The results of the confusion matrix of the two methods are presented in Table 4.

The accuracy of the model using and without data Augmentation is significantly different. For example, after testing with 46 different images, filtering with an accuracy of more than 60%, the model using data augmentation can recognize 42 images. In comparison, the origin model does not apply recognition of 38 images. Furthermore, most of the recognized images of the model using Data Augmentation have higher accuracy than the original model, for example, in Fig. 4. However, when examining images containing many objects, the original model can scan many objects, but the model using data augmentation is minimal, for example, in Fig. 6. This proves that data augmentation can improve the model's accuracy. The results of Precision and Recall of these two models are presented in Fig. 5. The calculation results of the two models are as follows in Table 4. The f1-score of the original model is about 94.05%, and the model using data augmentation is 94.83%. After the above comparison results, we can see that using data augmentation can improve the model's accuracy but limit its ability to recognize multiple objects in a single image.

Table 4. Confusion Matrix data without/using data augmentation

	Actual Class					
	ant	bear	bee	boa	buffalo	ACC < 60%
ant	0.92	0	0	0	0	0.08
bear	0	0.88	0	0	0	0.12
bee	0	0	0.55	0	0	0.45
boa	0	0	0	0.80	0	0.20
buffalo	0	0	0	0	0.88	0.12
	using data augmentation					
	ant	bear	bee	boa	buffalo	ACC < 60%
ant	0.92	0	0	0	0	0.08
bear	0	1	0	0	0	0
bee	0	0	0.92	0	0	0.08
boa	0	0	0	1	0	0
buffalo	0	0	0	0	0.63	0.37

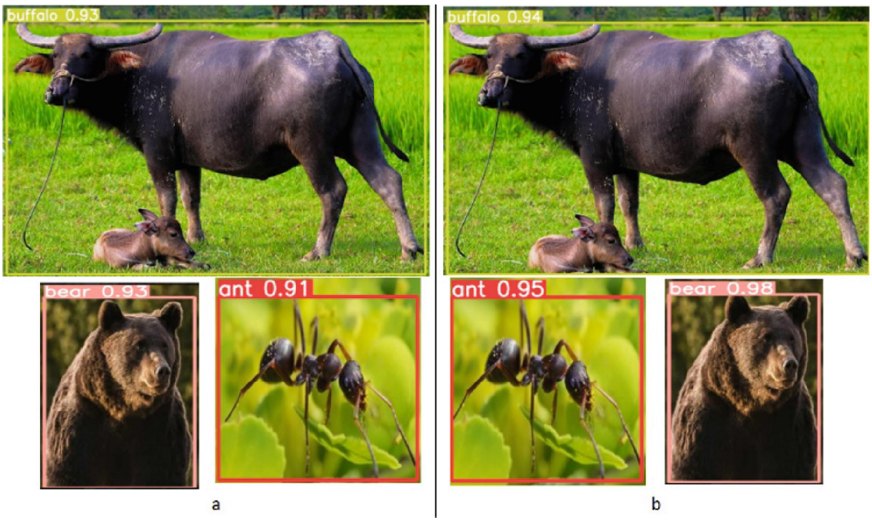

Fig. 4. Some illustration of objects recognized with bounding boxes: (a) without using data augmentation (b) using data augmentation

4.3 N-Object Detection Performance Comparison Between the Model Training Using the Architecture Starting from Scratch and the Model Starting from a Pre-trained Model Performing the Classification Tasks on the N-1 Objects

From the line charts below in Fig 7, the blue line represents training the model from the beginning. The orange line represents the training model from a previously trained model and continues to train another layer (model n+1). Comparing the object loss we see at the beginning, the orange line hits the threshold with a loss of 0.023585 but then starts to plummet, intersects and exceeds the blue line, and achieves a loss of 0.0091255. The lower the loss is, the higher the accuracy of the data and the higher the model will be, but that is the result that we have performed on the training set. Although we achieved results and saturation earlier, we are training a model from an existing model, which still has some limitations that we can see on the validation set on the correct chart. The orange line (object loss of the model) increases at the last epoch, leading to the detection of the affected object. However, because the model chosen to use is the model with the best epoch and because of the savings, If we have more time to train, then the use of the training method continues from a previous model selected to make the final product that the article is aiming for, which is the object recognition model.

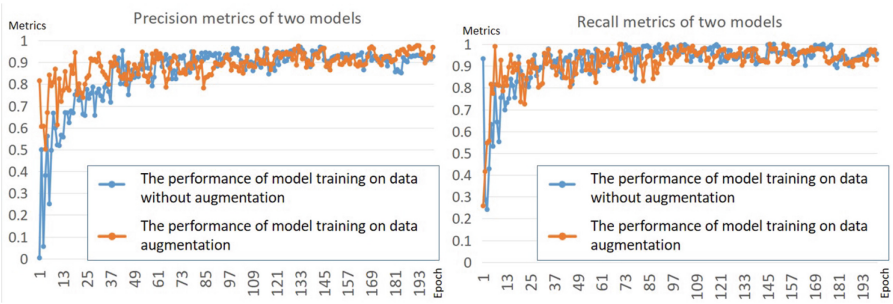

Fig. 5. Performance Comparison of the two models in the training phase

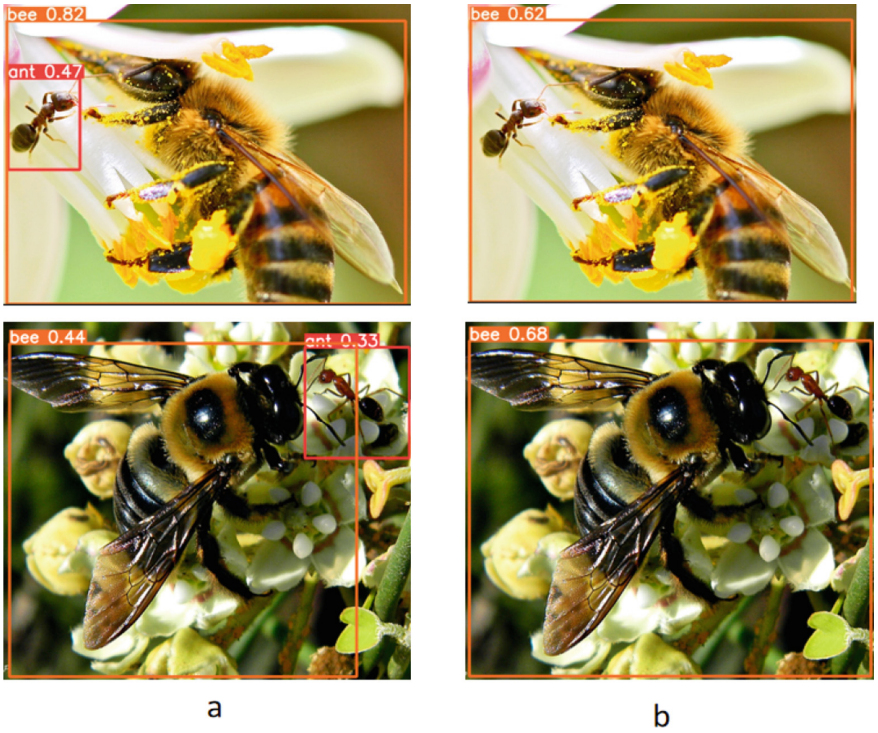

Fig. 6. Detection results on multiple objects in an image (a) without using data augmentation (b) using data augmentation

4.4 Develop an Application Using the Yolov5 Model to Support English Learning

The application will run on Android, developed in the Java programming language on Android Studio IDE. The application includes functions such as recognizing objects through images and videos to exporting a dictionary of that

object, and the application also provides a dictionary function to help users look up. The dictionary has been updated with a total of 151 words, so the model will be further developed to meet the previous dictionary to be able to update new words. The application interface is as shown in Fig. 8.

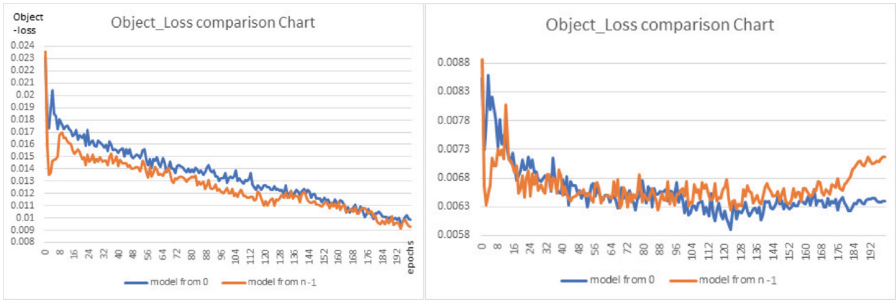

Fig. 7. Loss (y-axis) comparison through epochs (x-axis) between the model training from scratch (model from 0) and model training from a pre-trained on n-1 classes in the training set (left chart) and validation set (right chart)

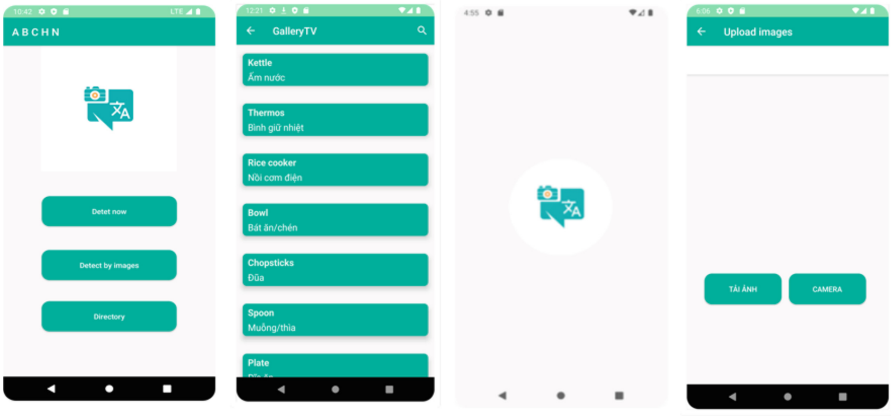

Fig. 8. Some illustrations of the application Interface

5 Conclusion

This study presented a workflow for object detection in images. We collected images from the internet and in Can Tho City, Vietnam. The data were labeled with Roboflow and then augmented by flipping original images to increase the

data size for the learning. As observed from the results, the flipping techniques can help to enhance the performance and push the model to converge quickly. In addition, we also see that the model training starting from a pre-trained model can converge faster.

This work is expected to be among the first steps to developing applications for supporting learning English by vocabulary extraction. Further study can collect more data to update and add more vocabulary.

Acknowledgement. This study is funded in part by the Can Tho University, Code: THS2022-15.

References

1. Liu, H., Aderon, C., Wagon, N., Liu, H., MacCall, S., Gan, Y.: Deep learning-based automatic player identification and logging in American football videos. arXiv preprint arXiv:2204.13809 (2022)
2. Zou, S., et al.: TOD-CNN: an effective convolutional neural network for tiny object detection in sperm videos. arXiv preprint arXiv:2204.08166 (2022)
3. Zhao, W., et al.: A survey of semen quality evaluation in microscopic videos using computer assisted sperm analysis. arXiv preprint arXiv:2202.07820 (2022)
4. Gu, Y., Liao, X., Qin, X.: YouTube-GDD: a challenging gun detection dataset with rich contextual information. arXiv preprint arXiv:2203.04129 (2022)
5. Yin, Q., et al.: Detecting and tracking small and dense moving objects in satellite videos: a benchmark. IEEE Trans. Geosci. Remote Sens. **60**, 1–18 (2022). https://doi.org/10.1109/TGRS.2021.3130436
6. Zhu, X., Dai, J., Yuan, L., Wei, Y.: Towards high performance video object detection. arXiv preprint arXiv:1711.11577 (2017)
7. Tang, P., Wang, C., Wang, X., Liu, W., Zeng, W., Wang, J.: Object detection in videos by high quality object linking. arXiv preprint arXiv:1801.09823 (2018)
8. He, F., Gao, N., Jia, J., Zhao, X., Huang, K.: QueryProp: object query propagation for high-performance video object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 1, pp. 834–842 (2022). https://doi.org/10.1609/aaai.v36i1.19965
9. Han, M., Wang, Y., Chang, X., Qiao, Y.: Mining inter-video proposal relations for video object detection (2020). https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123660426.pdf
10. Kolarova, S.T.V., et al.: Autonomous driving (2016). https://www.ifmo.de/files/publications_content/2016/ifmo_2016_Autonomous_Driving_2035_en.pdf
11. Advantech Co., Ltd.: The future of intelligent surveillance (2012). https://advcloudfiles.advantech.com/ecatalog/MyAdvantech/MyAdvantech_No_11_eng.pdf
12. Han, H., et al.: Real-time robust video object detection system against physical-world adversarial attacks. arXiv preprint arXiv:2208.09195 (2022)
13. Schofield, D., et al.: Chimpanzee face recognition from videos in the wild using deep learning. Sci. Adv. **5**(9), eaaw0736 (2019). https://www.science.org/doi/abs/10.1126/sciadv.aaw0736
14. Ardianto, S., Hang, H.M., Cheng, W.H.: Fast vehicle detection and tracking on fisheye traffic monitoring video using CNN and bounding box propagation. arXiv preprint arXiv:2207.01183 (2022), to be published in International Conference on Image Processing (ICIP) 2022, Bordeaux, France

15. Raskar, P.S., Shah, S.K.: Real time object-based video forgery detection using YOLO (V2) (2021). https://doi.org/10.1016/j.forsciint.2021.110979
16. Jiang, C., et al.: Object detection from UAV thermal infrared images and videos using YOLO models (2022). https://doi.org/10.1016/j.jag.2022.102912
17. Torresani, G.B.L., Shi, J.: Object detection in video with spatiotemporal sampling networks (2018). https://openaccess.thecvf.com/content_ECCV_2018/papers/Gedas_Bertasius_Object_Detection_in_ECCV_2018_paper.pdf
18. Deng, H., et al.: Object guided external memory network for video object detection (2019). https://ieeexplore.ieee.org/document/9011008
19. Oh, S.W., University, Y., Lee, J.Y., Research, A., Xu, N., Research, A., Kim, S.J., University, Y.: Video object segmentation using space-time memory networks (2019). https://openaccess.thecvf.com/content_ICCV_2019/papers/Oh_Video_Object_Segmentation_Using_Space-Time_Memory_Networks_ICCV_2019_paper.pdf
20. Fan, Q., Tang, C.K., Tai, Y.W.: Few-shot video object detection (2021). https://www.researchgate.net/publication/351278547_Few-Shot_Video_Object_Detection#pf9
21. Ultralytics: Ultralytics yolov5. https://github.com/ultralytics/yolov5. Accessed 27 Sep 2023