The Journal of Applied Science **Vol. 21 No. 2 [2022]: 245319**
วารสารวิทยาศาสตร์ประยุกต์ doi: 10.14416/j.appsci.2022.02.013
ISSN 1513-7805 (Printed in Thailand) ISSN 2586-9663 (Online)

Research Article

การเรียนรู้เชิงลึกสำหรับตรวจจับความผิดปกติจากภาพในวิดีโอการ์ตูนสำหรับเด็ก

Vision-based Deep Learning for Anomaly Detection from Video Cartoon for Kids

ธนานพ จันภักดี[1] และ ฐิติรัตน์ ศิริบวรรัตนกุล[2]*

Tananop Janpakdee [1] and Thitirat Siriborvornratanakul [2]*

[1,2] คณะสถิติประยุกต์ สถาบันบัณฑิตพัฒนบริหารศาสตร์ 148 หมู่ 3 ถนนเสรีไทย คลองจั่น บางกะปิ กรุงเทพฯ 10240 ประเทศไทย

[1,2] Graduate School of Applied Statistics, National Institute of Development Administration, 148 Serithai Road, Klong-Chan, Bangkapi, Bangkok, 10240, Thailand

***E-mail**: thitirat@as.nida.ac.th

Received: 01/09/2021;Revised: 29/06/2022; Accepted: 23/08/2022

บทคัดย่อ

จากผลสำรวจในปี พ.ศ. 2563 พบว่า เด็กไทย Generation Z มีการใช้งานสื่อโซเชียลมีเดียประเภท Youtube มากเป็นอันดับหนึ่ง ซึ่งปัญหาที่ตามมาคือวิดีโอสำหรับเด็กจำนวนหนึ่งบนแพลตฟอร์ม Youtube นั้นถูกสอดแทรกไว้ด้วยเนื้อหาที่ไม่เหมาะสมกับเด็ก ทำให้เด็กที่ดูเกิดอาการตื่นตกใจ หวาดกลัว หรือเกิดพฤติกรรมเลียนแบบที่ไม่พึงประสงค์ได้ อย่างไรก็ตาม การที่ผู้ปกครองจะมาตรวจสอบภาพทุกภาพในวิดีโอเพื่อหาสิ่งผิดปกติซึ่งอาจมีสอดแทรกอยู่เพียงเล็กน้อยในวิดีโอหนึ่ง ๆ นั้นเป็นเรื่องที่กินเวลามาก และแทบจะเป็นไปไม่ได้สำหรับจำนวนวิดีโอที่มีอยู่มากมายมหาศาลบน Youtube ด้วยเหตุนี้ผู้วิจัยจึงมีแนวคิดจะนำเทคนิคการวิเคราะห์ภาพด้วยการเรียนรู้เชิงลึกมาช่วยในการตรวจจับภาพเนื้อหาที่ไม่เหมาะสมในวิดีโอการ์ตูนสำหรับเด็กโดยอัตโนมัติ

งานวิจัยนี้มีจุดประสงค์เพื่อตรวจหาเฟรมภาพที่ผิดปกติจากวิดีโอการ์ตูนของเด็กโดยใช้เทคนิคการเรียนรู้เชิงลึก ซึ่งความผิดปกติที่เป็นเป้าหมายของงานนี้ ได้แก่ ภาพที่เป็นฉากเลือด และ ภาพที่มีอาวุธ (ปืนหรือมีด) ปรากฏอยู่ในภาพ โดยในส่วนของการตรวจจับฉากเลือดนั้นผู้วิจัยใช้เทคนิคถ่ายโอนความรู้จากโมเดลต้นแบบ ResNet50, VGG16 และ VGG19 เพื่อจำแนกภาพแต่ละภาพว่าเป็นหรือไม่เป็นฉากเลือด ผลการทดลองพบว่าโมเดลต้นแบบ VGG16 ให้ผลลัพธ์ดีที่สุด คือ ความแม่นยำเท่ากับ 0.84 และพื้นที่ใต้กราฟ ROC เท่ากับ 0.92 และในส่วนของการตรวจจับอาวุธผู้วิจัยใช้เทคนิคถ่ายโอนความรู้กับโมเดลต้นแบบสำหรับตรวจจับวัตถุ 2 ตัว ได้แก่ YOLOv3 และ YOLOv4 ผลการทดลองพบว่า YOLOv4 ให้ผลลัพธ์ที่ดีกว่าในการตรวจจับอาวุธมีดและปืนบนภาพการ์ตูน

คำสำคัญ: การเรียนรู้เชิงลึก การวิเคราะห์วิดีโอ การตรวจจับความผิดปกติ การจำแนกภาพ การถ่ายโอนการเรียนรู้

Abstract

According to the survey in 2020, Youtube is the most popular social media platform for Thai children in generation Z. Due to the overwhelming number of kid videos on Youtube, it is impossible for parents to check every single video and thoroughly look at every single image frame for any possible frame with inappropriate contents. To help ease this problem, this paper proposes using vision-based deep learning techniques to automatically detect frames with inappropriate visual content for children. In this work, our focus is on analyzing cartoon videos and the inappropriate content refers to blood scenes and weapons (gun and knife). To detect blood scenes, we use transfer learning and finetuning techniques with three pre-trained image classification backbone models—ResNet50, VGG16, and VGG19. Our experimental results show that VGG16 is the best that gives the highest accuracy of 0.84 and the highest area under ROC of 0.92. As for the part of weapon detection, we do transfer learning on two pre-trained object detection models, YOLOv3 and YOLOv4. Our results reveal that YOLOv4 is better at detecting guns and knives in cartoon videos.

Keywords: Deep Learning, Video analytics, Anomaly Detection, Image Classification, Transfer Learning

บทนำ

อ้างอิงจากผลสำรวจพฤติกรรมผู้ใช้อินเทอร์เน็ตในประเทศไทย ปี 2563 (ETDA, 2020) ของกลุ่มคน Generation Z (ผู้ที่มีอายุต่ำกว่า 20 ปี) พบว่าคนวัยนี้มีการใช้อินเทอร์เน็ตเฉลี่ย 12 ชั่วโมง 8 นาทีต่อวัน โดยกิจกรรมยอดฮิต 3 อันดับแรกได้แก่ โซเชียลมีเดีย (92.9%) ดูโทรทัศน์/ดูคลิป/ฟังเพลงออนไลน์ (85.7%) และ ค้นหาข้อมูลทางออนไลน์ (77.1%) โซเชียลมีเดียยอดฮิตที่คนวัยนี้ชื่นชอบ ได้แก่ Youtube (98.6%), Facebook (97.3%) และ LINE (93.4%) ตามลำดับ ในบริบทของเด็กและเยาวชน การที่ผู้ปกครองปล่อยให้เด็กเข้าถึงสื่อออนไลน์อย่างอิสระ สามารถก่อให้เกิดผลเสียกับเด็กได้หลายประการ เช่น จากการศึกษาอิทธิพลของสื่อที่ส่งผลต่อพฤติกรรมความรุนแรงของเด็กและเยาวชน จังหวัดเชียงใหม่ (Ratchatathanarat & Wongcharoen, 2021) พบว่า ปัจจัยแวดล้อมด้านภาพและเนื้อหาของสื่อจะส่งผลต่อพฤติกรรมความรุนแรงของเด็กและเยาวชนสูงที่สุด และตัวแปรด้านสื่อ ด้านสิ่งเร้าของสื่อ ด้านการรับรู้ และด้านการเรียนรู้นั้นมีอิทธิพลทั้งทางตรงและทางอ้อมต่อพฤติกรรมความรุนแรงของเด็กและเยาวชน นอกจากนี้ในผลการศึกษาของ (Tobbi, 2018) ก็พบว่าโซเชียลมีเดียสามารถส่งผลต่อสุขภาพจิตของเด็กได้ เช่น ภาวะซึมเศร้า ความวิตกกังวล และการนอนหลับไม่ดี อย่างไรก็ตาม ด้วยปริมาณและความหลากหลายอันมหาศาลของข้อมูลบนโซเชียลมีเดียทำให้เจ้าของแพลตฟอร์มโซเชียลมีเดียต่าง ๆ ไม่สามารถจะคัดกรองหรือรับรองความถูกต้องของข้อมูลได้ทั้งหมด จึงเป็นหน้าที่ของผู้รับสื่อเองที่ต้องอาศัยประสบการณ์และความสามารถในการคิดวิเคราะห์ของตนเพื่อเลือกรับเฉพาะข้อมูลที่ถูกต้องและมีประโยชน์ แต่ในบริบทของเด็กนั้นการพินิจพิเคราะห์เหล่านี้ยังจำเป็นต้องมีผู้ปกครองคอยช่วยชี้แนะอยู่ข้าง ๆ ไม่เช่นนั้นเด็กอาจเกิดการเรียนรู้และเลียนแบบพฤติกรรมไม่เหมาะสมซึ่งถูกสอดแทรกเข้ามาในเนื้อหาของสื่อโซเชียลมีเดียได้

อ้างอิงจากประกาศของสำนักงานคณะกรรมการกิจการกระจายเสียง กิจการโทรทัศน์ และกิจการโทรคมนาคมแห่งชาติ (กสทช.) เรื่องหลักเกณฑ์การจัดทำผังรายการสำหรับการให้บริการกระจายเสียงหรือโทรทัศน์ พ.ศ. 2556 ได้กำหนดความเหมาะสมของเนื้อหาสำหรับผู้ชมในช่วงอายุ 6–12 ปีไว้ว่าไม่ควรมีเนื้อหาที่มีความรุนแรง และเรื่องทางเพศ โดยความรุนแรงประกอบไปด้วย (1) การแสดงพฤติกรรมที่ไม่เหมาะสมที่ก่อให้เกิดความรุนแรงต่อจิตใจของผู้ชม (2) การใช้ความรุนแรงต่อตนเอง บุคคลอื่น สิ่งมีชีวิต และวัตถุ (3) การใช้อาวุธ ยาเสพติด การแสดงภาพที่กระทำความผิดในรูปแบบต่างๆ ที่ขัดต่อศีลธรรม

(4) การนำเสนอที่ก่อให้เกิดการอคติ การเลือกปฏิบัติที่นำมาซึ่งการต่อต้านหรือล่วงละเมิดบุคคล รวมถึงการส่งเสริม การสร้างทัศนคติที่ทำให้เกิดการต่อต้าน การดูหมิ่นเหยียดหยาม นอกจากนี้หากอ้างอิงจากงานของ (Government Gazette, 2013) ยังมีการนิยามความรุนแรงที่รวมถึงพฤติกรรมที่นำไปสู่อันตราย เช่น การพยายามทำร้ายตัวเองและผู้อื่น การฆ่าตัวตาย การเลียนแบบกีฬาที่เป็นการต่อสู้ เป็นต้น

ในบริบทของสื่อสำหรับเด็กและเยาวชน การตรวจสอบความรุนแรงที่ถูกสอดแทรกอยู่ในข้อมูลบนโซเชียลมีเดียจะยิ่งทวีความยุ่งยากมากขึ้น เมื่อข้อมูลดังกล่าวอยู่ในรูปของวิดีโอซึ่งมีจำนวนเฟรมของภาพในวิดีโอมากมาย ยากที่ผู้ปกครองจะช่วยตรวจสอบภาพได้ครบหมดทุกเฟรม เป็นเหตุให้มีข่าวคราวอยู่เนือง ๆ ถึงเหตุการณ์ที่มีวิดีโอการ์ตูนสำหรับเด็กถูกผู้ไม่ประสงค์ดีสอดไส้เนื้อหาที่ไม่เหมาะสมเข้ามา จากปัญหาที่กล่าวมานี้ผู้วิจัยจึงต้องการนำปัญญาประดิษฐ์ (Artificial Intelligence) ในแขนงของการวิเคราะห์ภาพ (Image Analytics) มาพัฒนาเป็นระบบอัตโนมัติสำหรับช่วยตรวจหาเฟรมภาพที่ผิดปกติ (Anomaly Detection) ในวิดีโอการ์ตูนสำหรับเด็กอายุ 6–12 ปี โดยระบบต้นแบบปัจจุบันจะมุ่งเน้นไปที่การใช้เทคนิคการจำแนกภาพ (Image Classification) สำหรับช่วยตรวจจับฉากเลือดในเฟรมภาพ และเทคนิคการตรวจจับวัตถุ (Object Detection) สำหรับช่วยตรวจจับอาวุธ อันได้แก่ มีดและปืน

ปัจจุบันเทคนิคการเรียนรู้เชิงลึก (Deep Learning) ถูกนำมาประยุกต์ใช้ในการวิเคราะห์ภาพอย่างแพร่หลาย เช่น งานวิจัยของ (Nikita et al., 2021) ทำการรวบรวมภาพมาทั้งหมด 8,000 ภาพจากการ์ตูนเรื่อง Tom and Jerry และใช้เทคนิคถ่ายโอนความรู้ (Transfer Learning) บนโมเดลต้นแบบ ResNet50, MobileNetV2, InceptionV3 และ VGG16 เพื่อแยกแยะภาพแต่ละภาพว่าเป็นอารมณ์อะไร ผลการศึกษาพบว่า VGG16 ให้ผลลัพธ์ดีที่สุด หรือในงานวิจัยของ (Mohan, 2020) ที่ศึกษาการจำแนกภาพออกเป็น 2 กลุ่มระหว่าง Normal และ Pneumonia โดยใช้ข้อมูลภาพสำหรับฝึกอบรม 5,856 ภาพ และใช้โมเดลต้นแบบ VGG16, VGG19, ResNet50, InceptionV3, Xception และ DenseNet ในการจำแนก ผลการศึกษาพบว่า Xception และ VGG16 ให้ประสิทธิภาพในการทำงานที่ดี ไม่เพียงเท่านั้นยังมีงานวิจัยของ (Noreen et al., 2019) ที่จำแนกฉากผิดปกติของการ์ตูนโดยใช้ MobileNet และแบ่งความผิดปกติออกเป็น Fight, Gunshot, Screaming, Blood, Fire และ Explosion ทำการทดลองด้วยข้อมูลภาพประเภทละ 500 ภาพ ในส่วนของงานวิจัย (Rashid et al., 2019) ได้ทดลองจำแนกวิดีโอการ์ตูนออกเป็น Original Video, Fake Explicit และ Fake Violence โดยใช้โมเดล ResNet, VGG16, VGG19 และ Google LeNet ผลการศึกษาพบว่า VGG19 ให้ผลลัพธ์ดีที่สุด ในส่วนของการตรวจจับอาวุธนั้นอ้างอิงจากงานวิจัย (Sanam et al., 2021) ที่ใช้โมเดล YOLOv2 และYOLOv3 ในการตรวจจับอาวุธ พบว่าประสิทธิภาพของ YOLOv3 ให้ผลดีที่สุด ส่วนงานวิจัยของ (Lamoonmorn & Sucontphunt, 2019) ที่ศึกษาการตรวจจับหมวกนิรภัยและการใช้อาวุธปืนเพื่อเตือนภัยเหตุโจรกรรม โดยใช้โมเดล YOLOv3 และ YOLOv4 ก็พบว่าโมเดล YOLOv4 ให้ประสิทธิภาพดีที่สุดในการตรวจจับ

วิธีการดำเนินงานวิจัย

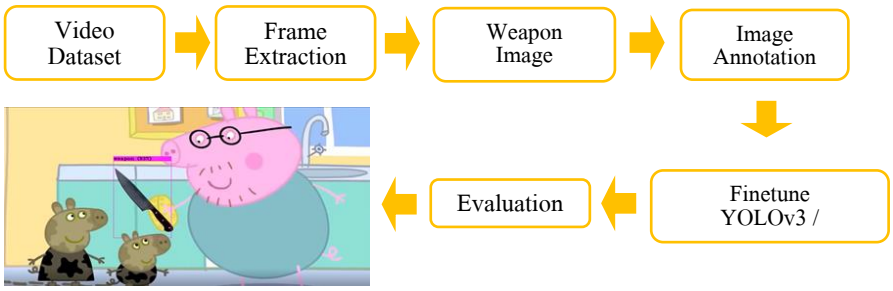
ในการดำเนินงานผู้วิจัยใช้ภาษาไพธอนและใช้แพลตฟอร์ม Google Colab โดยมีไลบรารีและโปรแกรมที่เกี่ยวข้องกับการพัฒนา ได้แก่ Tensorflow 2.5.0 (Keras API), OpenCV 4.1.2 สำหรับเขียนโครงสร้างในการจำแนกภาพ, Youtube-dl 2021.01.24.1 สำหรับดาวน์โหลดวิดีโอจาก Youtube, Windows Movie Maker 2012 สำหรับตัดต่อวิดีโอเพื่อทำข้อมูลวิดีโอทดสอบ และ FastImageResizer 0.98 สำหรับลดขนาดของภาพ

1. ภาพรวมของระบบ (System Overview)

ขั้นตอนของการทำงานเพื่อจำแนกเฟรมภาพว่าเป็นฉากเลือดหรือไม่นั้นแสดงในรูปที่ 1 โดยผู้วิจัยทำการรวบรวมวิดีโอการ์ตูน สกัดเฟรมภาพออกจากวิดีโอ จากนั้นจึงแยกเฟรมภาพที่เหมาะสม (ไม่มีเลือด) และไม่เหมาะสม (มีเลือด) ไว้คนละ

โฟลเดอร์ เมื่อเตรียมข้อมูลเสร็จแล้วผู้วิจัยก็นำภาพทั้งหมดไปฝึกสอนโมเดล ResNet50, VGG16 และ VGG19 ทำการประเมินโมเดล และนำโมเดลที่ให้ผลลัพธ์ดีที่สุดมาทำการทดสอบกับวิดีโอการ์ตูนที่เตรียมไว้อีกชุดหนึ่ง ในส่วนของขั้นตอนดำเนินงานเพื่อตรวจจับอาวุธจากวิดีโอการ์ตูนนั้นแสดงดังรูปที่ 2 กล่าวคือ ผู้วิจัยนำเฟรมภาพที่สกัดจากวิดีโอการ์ตูนมาเลือกเฉพาะภาพที่มีอาวุธแยกออกมา จากนั้นก็ทำ Image Annotation และนำข้อมูลที่ได้ไปใช้ฝึกสอนโมเดล YOLOv3 และ YOLOv4 สุดท้ายก็เลือกโมเดลที่ให้ผลลัพธ์ดีที่สุดมาทำการทดสอบกับวิดีโอการ์ตูนที่เตรียมไว้อีกชุดหนึ่ง

รูปที่ 1 แผนภาพการดำเนินงานของการจำแนกภาพการ์ตูนฉากที่มีเลือด (เครดิตภาพการ์ตูนระบุในหัวข้อ Data Collection)



รูปที่ 2 แผนภาพการดำเนินงานของการตรวจจับอาวุธจากวิดีโอการ์ตูน (เครดิตภาพการ์ตูนระบุในหัวข้อ Data Collection)

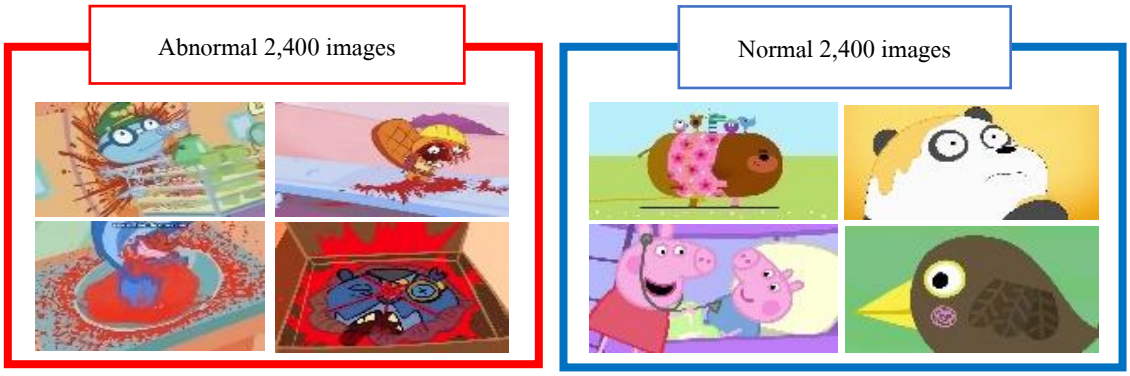
2. การเก็บรวบรวมข้อมูล (Data Collection)

ในงานนี้ผู้วิจัยทำการรวบรวมวิดีโอการ์ตูนจาก Youtube เป็นจำนวนทั้งหมด 80 วิดีโอ ได้แก่ Peppa Pig, Little Princess, HEY DUGGEE, Granny's Fairy Tales, Simon, Carl's Car Wash, Kid e Cats, Ben and Holly's Little Kingdom, Family Guy, Looney Tunes, Happy Tree Friends, OGGY, We Bare Bears, Tom & Jerry เป็นต้น เวลาของวิดีโอรวมทั้งหมดคือ 9 ชั่วโมง 11 นาที 58 วินาที นอกจากนี้ในส่วนของภาพอาวุธปืนและมีดผู้วิจัยมีการใช้ข้อมูลภาพจาก ARI-DaSCI (https://github.com/ari-dasci/OD-WeaponDetection) มาเสริมเพิ่มเติม

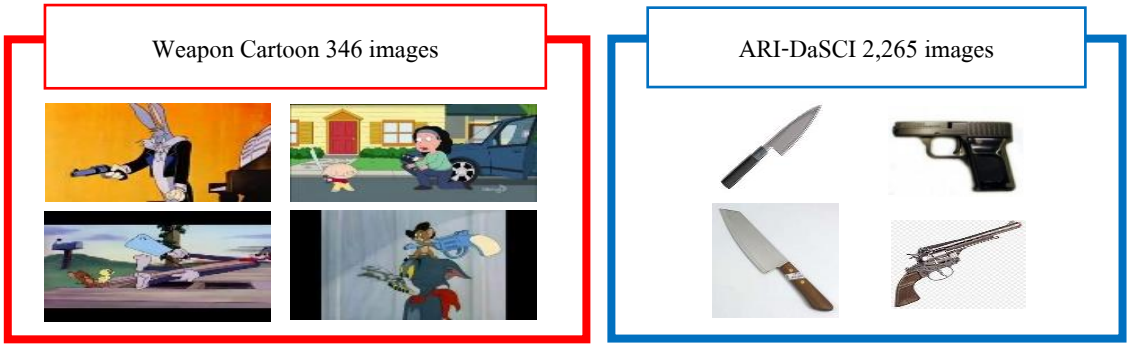
3. การจัดเตรียมข้อมูล (Data Preparation)

สำหรับการจำแนกภาพ (Image Classification) ออกเป็นเหมาะสม (Normal: ไม่มีเลือด) และไม่เหมาะสม (Abnormal: ฉากเลือด) ดังตัวอย่างภาพการ์ตูนในรูปที่ 3 นั้น ผู้วิจัยทำการสกัดเฟรมภาพจากวิดีโอที่รวบรวมมาโดยสกัดออกมาทุกๆ 0.5 วินาทีได้ภาพมาทั้งหมด 66,437 ภาพ จากนั้นก็นำภาพที่ได้มาแยกออกเป็นเฟรมภาพการ์ตูนที่เหมาะสมและไม่เหมาะสมได้จำนวน 64,037 ภาพ (96.39%) และ 2,400 ภาพ (3.61%) ตามลำดับ เนื่องจากปริมาณข้อมูลภาพการ์ตูนฉากเหมาะสมมีจำนวนมากกว่าฉากไม่เหมาะสมอย่างมีนัยสำคัญ ผู้วิจัยจึงทำการลดปริมาณข้อมูล (Undersampling) ภาพการ์ตูนฉากเหมาะสมให้เหลือเท่ากับปริมาณข้อมูลฉากไม่เหมาะสม อีกทั้งยังทำการลดขนาด (Resize) ภาพทั้งหมดเพื่อประหยัดเวลาในการอัปโหลดข้อมูลภาพไปใช้ใน Google Colab ด้วย ข้อมูลภาพทั้งหมดนี้ถูกนำมาแบ่งออกเป็น 2 ส่วน ได้แก่ ข้อมูลสำหรับฝึกอบรม (Train Set) จำนวน 1,920 ภาพ (80%) และข้อมูลสำหรับทดสอบ (Test Set) จำนวน 480 ภาพ (20%)

ในส่วนของการตรวจจับอาวุธ (Weapon Detection) นั้น ผู้วิจัยทำการแยกเฉพาะภาพการ์ตูนที่มีอาวุธออกมาและนำมารวมกับข้อมูลภาพอาวุธประเภทปืนและมีดจากชุดข้อมูลภาพ ARI-DaSCI ดังตัวอย่างในรูปที่ 4 ซึ่งจำนวนของภาพอาวุธที่ได้จากวิดีโอการ์ตูนและจาก ARI-DaSCI คือ 346 ภาพ (13.25%) และ 2,265 ภาพ (86.75%) ตามลำดับ โดยภาพการ์ตูนที่มีอาวุธ 346 ภาพนั้นผู้วิจัยได้คัดเลือกออกมาจากเฟรมภาพในวิดีโอทั้ง 66,437 ภาพด้วยตนเอง เมื่อรวมทั้งหมดแล้วจะได้ภาพที่ใช้ในงานส่วนนี้ทั้งหมด 2,611 ภาพ ประกอบด้วยภาพปืน 1,910 ภาพ และภาพมีด 701 ภาพ ซึ่งผู้วิจัยทำการแบ่งข้อมูลทั้งหมดนี้ออกเป็นข้อมูลสำหรับฝึกอบรม 80% (ปืน 1,528 ภาพ และมีด 561 ภาพ) และข้อมูลสำหรับทดสอบ 20% (ปืน 382 ภาพ และมีด 140 ภาพ)



รูปที่ 3 ตัวอย่างภาพการ์ตูนที่ไม่เหมาะสม (Abnormal) และเหมาะสม (Normal) สำหรับใช้ในงานการจำแนกภาพ (เครดิตภาพการ์ตูนระบุในหัวข้อ Data Collection)



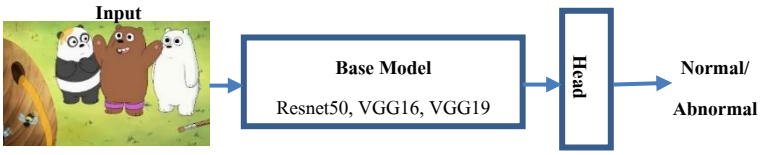
รูปที่ 4 ตัวอย่างข้อมูลภาพสำหรับใช้ในงานการตรวจจับอาวุธปืนและมีด (เครดิตภาพระบุในหัวข้อ Data Collection)

4. การจำแนกภาพ (Image Classification)

เนื่องจากข้อมูลภาพที่เหมาะสม (Normal: ไม่มีเลือด) และไม่เหมาะสม (Abnormal: มีเลือด) ทั้ง 2 ประเภทนั้นมีอยู่เพียงประเภทละ 2,400 ภาพเท่านั้น ซึ่งอาจไม่เพียงพอต่อการฝึกอบรมโมเดลการเรียนรู้เชิงลึกและอาจส่งผลต่อความแม่นยำของโมเดล ดังนั้นสำหรับข้อมูลชุดฝึกอบรมผู้วิจัยจึงใช้เทคนิค Data Augmentation เสริมเข้ามา โดยเทคนิคนี้นอกจากจะช่วยเพิ่มจำนวนข้อมูลฝึกอบรมให้มากขึ้นแล้วยังช่วยป้องกันปัญหา Overfit ของโมเดลการเรียนรู้เชิงลึกได้ด้วย ในงานวิจัยนี้ผู้วิจัยเลือกใช้เทคนิค Data Augmentation ได้แก่ การหมุนภาพ (Rotation) การปรับความสว่างของภาพ (Brightness) การพลิกภาพตามแนวนอน (Horizontal) การพลิกภาพตามแนวตั้ง (Vertical) โดยมีการกำหนดค่าที่ใช้ดังตารางที่ 1

ตารางที่ 1 การตั้งค่าของพารามิเตอร์ในไลบรารี Keras สำหรับทำ Data Augmentation ในงานวิจัยนี้

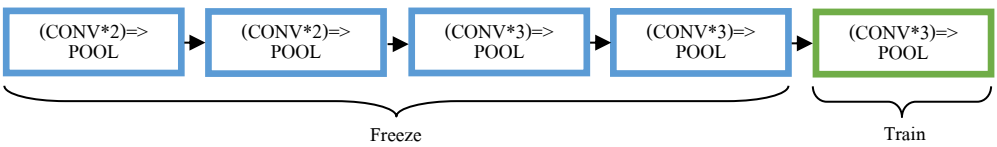
Type of Augmentation	Values
Rotation	45°
Brightness	0.2 to 1
Horizontal	Yes
Vertical	Yes



รูปที่ 5 โครงสร้างทั่วไปของโมเดลการเรียนรู้เชิงลึกที่ใช้เทคนิค Transfer Learning สำหรับจำแนกภาพออกเป็นฉากเหมาะสม (Normal) และไม่เหมาะสม (Abnormal) (เครดิตภาพการ์ตูนระบุในหัวข้อ Data Collection)

สำหรับการฝึกสอนโมเดลนั้นผู้วิจัยนำโมเดลการเรียนรู้เชิงลึก (Base Model) ได้แก่ ResNet50, VGG16 และ VGG19 ที่เคยถูกฝึกสอนไว้แล้วกับชุดข้อมูลภาพ ImageNet มาต่อยอดทำการทดลอง Transfer Learning และ Finetuning หลาย ๆ แบบ โดยจะมีการเพิ่มโมเดลส่วนต่อยอด (Head Model) เข้าไป และนำภาพการ์ตูนที่ผู้วิจัยเตรียมไว้มาทำการฝึกสอน เพื่อให้โมเดลเหล่านี้สามารถนำความรู้เดิมจากชุดข้อมูล ImageNet มาต่อยอดและปรับใช้ให้เข้ากับชุดข้อมูลภาพการ์ตูนในงานวิจัยนี้ได้ รูปที่ 5 แสดงแนวคิดของการทำ Transfer Learning ในงานวิจัยนี้

ในงานวิจัยนี้ผู้วิจัยได้ทำการทดลองหลายแบบเพื่อหาส่วนผสมของการฝึกสอนโมเดลที่จะได้ผลการจำแนกภาพที่ดีที่สุด เนื่องจากแต่ละ transfer learning ที่นำมาทำการทดลอง มีเลเยอร์ภายในไม่เท่ากัน เพื่อให้เกิดความเท่าเทียมในการเปรียบเทียบ ผู้วิจัยจึงใช้ classifier head และ setting ต่าง ๆ เหมือนกันทุกการทดลองเพื่อใช้เปรียบเทียบประสิทธิภาพ แต่สิ่งที่เปลี่ยนคือมีการทดลองทั้งการ freeze ทุกเลเยอร์ในส่วน Base Model และการ freeze แค่บางเลเยอร์ในส่วน Base Model ได้แก่ ในการทดลองที่ 1–3 ผู้วิจัยทำการ freeze ทุกเลเยอร์ในส่วนของ Base Model (ResNet50, VGG16 และ VGG19) เอาไว้และทำการฝึกสอนปรับปรุงเฉพาะน้ำหนักของส่วน Head Model (classifier head) เท่านั้น ในการทดลองที่ 4–5 ผู้วิจัยทำการ freeze ทุกเลเยอร์ของ VGG16 และVGG19 ยกเว้นเพียง 4 เลเยอร์สุดท้าย (ดังแสดงในรูปที่ 6) โดยการปรับแต่งอ้างอิงจากงานวิจัย (Burugupalli, 2020) ในการทดลองที่ 6 ผู้วิจัยทำการ freeze เฉพาะ 10 เลเยอร์แรกของ ResNet50 และต่อมาในการทดลองที่ 7-8 ผู้วิจัยได้ทดลองกำหนดส่วนของ Head Model เป็น 2 รูปแบบสรุปดังตารางที่ 2 ทั้งนี้ ในส่วนของการตั้งค่าพารามิเตอร์ที่เกี่ยวข้องกับการฝึกสอนนั้น ผู้วิจัยเลือกใช้ Binary Cross Entropy Loss และการทดลองที่ 9-10 ผู้วิจัยก็ได้ทำการทดลองเปรียบเทียบผลระหว่างการใช้ Optimizer ในการฝึกสอนสองแบบได้แก่ Stochastic Gradient Descent (SGD) และ Adaptive Moment Estimation (Adam) โดยมีการกำหนดค่าพารามิเตอร์ที่เกี่ยวข้องกับ Optimizer สรุปดังตารางที่ 3 ซึ่งการ finetuning ตัวแปร Dense ด้วยค่า 512 และตัวแปร Optimizer ด้วยค่า 0.0001 ปรับแต่งอ้างอิงจากงานวิจัย (Tahir et al., 2019)



รูปที่ 6 ตัวอย่างการ freeze เฉพาะ 4 เลเยอร์สุดท้ายของ VGG16 ในการทดลองที่ 4 ของงานส่วนการจำแนกภาพ

ตารางที่ 2 รายละเอียดของ Head Model สองรูปแบบสำหรับการทดลองที่ 7-8 ของงานการจำแนกภาพ

Type of Layers	Head Model	
	(1)	(2)
AveragePooling2D	✓	✓
Flatten Layer	✓	✓
Dense	512	64
Activation	Relu	Relu
Dropout	0.50	0.25
Dense	512	64
Activation	Relu	Relu
Dense	2	2
Activation	SoftMax	SoftMax

ตารางที่ 3 ค่าพารามิเตอร์ในส่วนของ Optimizer สองแบบ สำหรับการทดลองที่ 9-10 ของงานการจำแนกภาพ

Type of Parameter	Optimizer	
	SGD	Adam
Learning Rate	0.0001	0.0001
Momentum	0.9	-
Decay	0.0001/300	0.0001/300

5. การตรวจจับอาวุธ (Weapon Detection)

ผู้วิจัยนำข้อมูลภาพที่เตรียมไว้ทุกภาพมาทำการลดขนาด (Resize) และนำภาพดังกล่าวมาทำการตีกรอบสี่เหลี่ยม (bounding box) เพื่อระบุตำแหน่งของอาวุธที่สนใจในภาพโดยใช้โปรแกรม labelImg (https://github.com/tzutalin/labelImg) ดังตัวอย่างของการลากกรอบสี่เหลี่ยมในรูปที่ 7 นอกจากนี้เพื่อให้การฝึกอบรมไม่ใช้เวลาประมวลผลนานเกินไป ผู้วิจัยจึงเลือกใช้เทคนิค Transfer Learning และ Finetuning อีกครั้งโดยนำโมเดล YOLOv3 และ YOLOv4 ที่เคยถูกฝึกอบรมไว้แล้วบนชุดข้อมูลอื่น มาทำการฝึกสอนต่อยอดบนชุดข้อมูลภาพที่ผู้วิจัยได้รวบรวมมาเอง ทั้งนี้พารามิเตอร์ที่ตั้งไว้สำหรับ YOLOv3 และ YOLOv4 นั้นคล้ายคลึงกัน คือ batch = 64, subdivisions = 32, max_batches = 4000, steps = 3200 และ 3600, filters = 21, random = 0 ในส่วนของพารามิเตอร์อื่น ๆ จะใช้ค่าตั้งต้นจากตัวต้นแบบในเฟรมเวิร์ก darknet YOLOv3 และdarknet YOLOv4

รูปที่ 7 ตัวอย่างการตีกรอบสี่เหลี่ยมเพื่อกำหนดตำแหน่งวัตถุที่สนใจด้วยโปรแกรม labelImg (เครดิตภาพการ์ตูนระบุในหัวข้อ Data Collection)

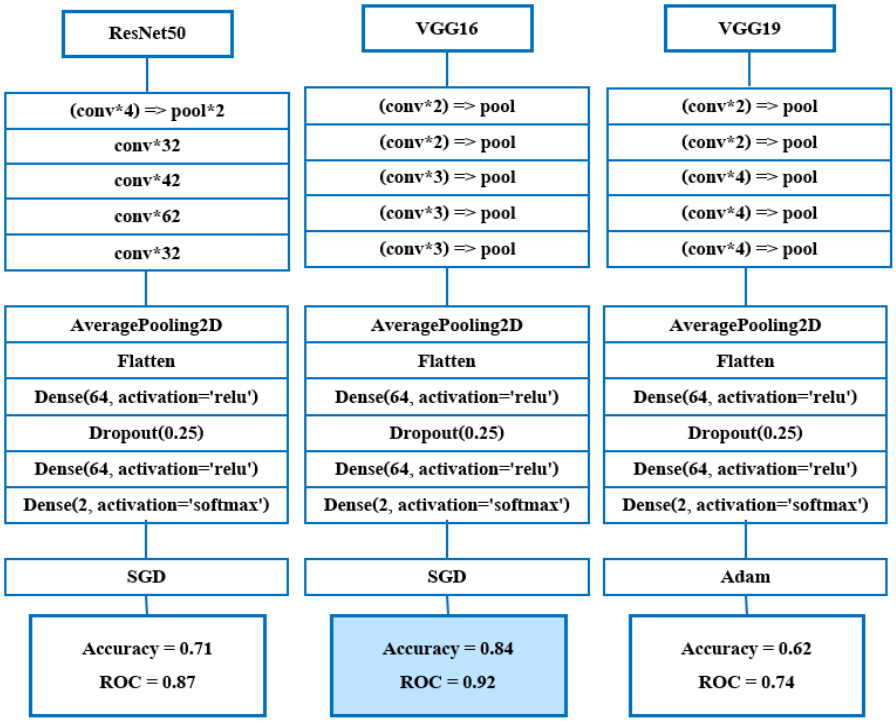
ผลการทดลอง

1. ผลลัพธ์ของการฝึกอบรม: การจำแนกภาพการ์ตูน

ในหัวข้อนี้ผู้วิจัยทำการประเมินโมเดลจำแนกภาพการ์ตูนที่ถูกฝึกอบรมเสร็จแล้วด้วยการพิจารณาจากค่า Accuracy และค่าพื้นที่ใต้กราฟของ Receiver Operating Characteristic (ROC) Curve เพื่อเลือกโมเดลที่ดีที่สุดออกมา ทั้งนี้ในการพิจารณา Classification Report ผู้วิจัยจะให้น้ำหนักความสำคัญในส่วน Recall มากกว่า เนื่องจากถ้ามีฉากไม่เหมาะสมหลุดรอดให้เด็กรับชมจะทำให้เกิดผลเสียมากกว่า ทั้งนี้ทรัพยากรที่ผู้วิจัยใช้ในการฝึกอบรมระบบส่วนจำแนกภาพคือ Google Colab Pro (High-RAM และ GPU: Tesla P100) ใช้ข้อมูลชุดฝึกอบรม (Train Set) ในการฝึกอบรมแต่ละโมเดลนานประมาณ 3 ชั่วโมงต่อหนึ่งโมเดล อบรมแต่ละโมเดลเป็นจำนวน 300 epochs

รูปที่ 8 แสดงผลลัพธ์ที่ดีที่สุดที่ได้จากการประเมิน Base Model ทั้งสามแบบ (ResNet50, VGG16, VGG19) บนข้อมูลชุดทดสอบ (Test Set) จากรูปจะเห็นว่าโมเดล ResNet50 ที่ดีที่สุด (จากผลการทดลองที่ 6+8+9) ให้ Accuracy เท่ากับ 0.71 และ ROC เท่ากับ 0.87 ส่วนโมเดล VGG16 ที่ดีที่สุด (จากผลการทดลองที่ 4+8+9) ให้ Accuracy เท่ากับ 0.84 และ ROC เท่ากับ 0.92 และสุดท้ายโมเดล VGG19 ที่ดีที่สุด (จากผลการทดลองที่ 5+8+10) ให้ Accuracy เท่ากับ 0.62 และ ROC เท่ากับ 0.74 จากผลการทดลองนี้สรุปได้ว่า VGG16 คือโมเดลที่ให้ผลลัพธ์ดีที่สุดในการจำแนกระหว่างภาพการ์ตูนที่เหมาะสม (Normal: ไม่มีเลือด) และไม่เหมาะสม (Abnormal: มีเลือด)

การทดลองที่ 1–3 ซึ่งผู้วิจัยทำการ freeze ทุกเลเยอร์ในส่วนของ Base Model นั้น ให้ผลลัพธ์ไม่ดีเท่ากับการเลือก freeze เพียงบางเลเยอร์ เช่น โมเดล ResNet50 (จากผลการทดลองที่ 1 และ 8 และ 9) ให้ Accuracy เท่ากับ 0.60 และ ROC เท่ากับ 0.68 ส่วนโมเดล VGG16 (จากผลการทดลองที่ 2 และ 7 และ 9) ให้ Accuracy เท่ากับ 0.49 และ ROC เท่ากับ 0.13 และสุดท้ายโมเดล VGG19 (จากผลการทดลองที่ 3 และ 7 และ 9) ให้ Accuracy เท่ากับ 0.48 และ ROC เท่ากับ 0.49

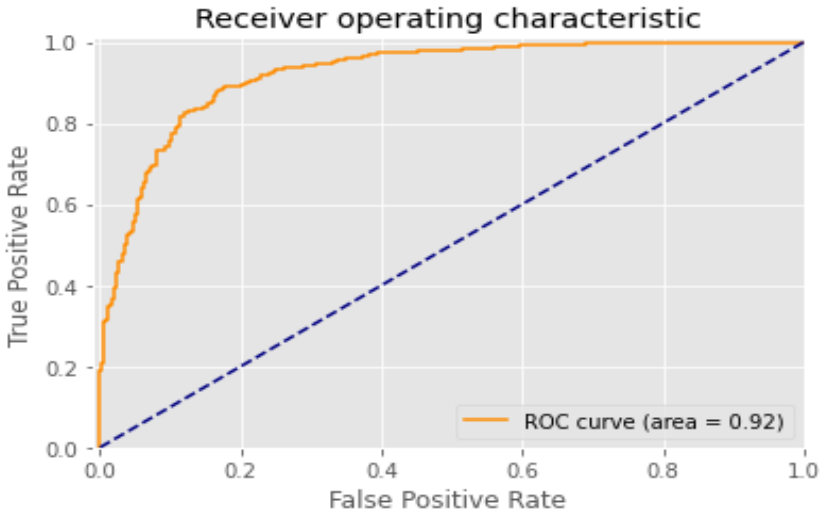


รูปที่ 8 สรุปผลลัพธ์การประเมินบนข้อมูลชุดทดสอบที่ดีที่สุด สำหรับ Base Model แต่ละแบบที่ใช้ในงานนี้

นอกจากนี้ผู้วิจัยยังนำโมเดล VGG16 ตัวที่ดีที่สุด (ดังข้อสรุปจากย่อหน้าก่อน) มาทำการประเมินประสิทธิภาพโดยละเอียดเพิ่มเติมอีกดังแสดงในรูปที่ 9 และตารางที่ 4 โดยผลการประเมินได้พื้นที่ใต้กราฟ ROC เท่ากับ 0.92 และได้รายงานการจำแนก (Classification Report) คือ Precision (abnormal) เท่ากับ 0.89, Precision (normal) เท่ากับ 0.80, Recall (abnormal) เท่ากับ 0.78 และ Recall (normal) เท่ากับ 0.91, F1-Score (abnormal) เท่ากับ 0.83, F1-Score (normal) เท่ากับ 0.85 เมื่อ Support คือ จำนวนข้อมูลภาพที่นำมาใช้ในการทดสอบนั้น ๆ

ตารางที่ 4 Classification Report ของโมเดล VGG16 ที่ดีที่สุด ทำการประเมินบนชุดข้อมูลภาพทดสอบ (Test Set)

	Precision	**Recall**	**F1-Score**	**Support**
Abnormal	0.89	0.78	0.83	480
Normal	0.80	0.91	0.85	480
Accuracy			0.84	960



รูปที่ 9 ROC Curve ของโมเดล VGG16 ที่ดีที่สุด แกน X คือ False Positive Rate และแกน Y คือ True Positive Rate

2. ผลลัพธ์ของการฝึกอบรม: การตรวจจับอาวุธ

สำหรับการตรวจจับอาวุธจากภาพในวิดีโอการ์ตูนสำหรับเด็ก ผู้วิจัยทำการฝึกอบรมชุดข้อมูลภาพที่เตรียมไว้บน Google Colab ซึ่งใช้ GPU: Tesla K80 โดยทำการฝึกอบรมวนซ้ำ 3 รอบ การฝึกอบรมซ้ำจำนวนหลายรอบนี้มีจุดประสงค์เพื่อเฉลี่ยประสิทธิภาพส่วนหนึ่งของแบบจำลองที่ถูกตั้งต้นด้วยค่าน้ำหนักสุ่มที่แตกต่างกัน ซึ่งน้ำหนักสุ่มดังกล่าวนั้นมีความแตกต่างที่เป็นไปได้ระดับอนันต์ ไม่ทราบจำนวนที่แน่นอน ในทางปฏิบัติจริงจึงไม่ค่อยพบการฝึกอบรมซ้ำมาก ๆ รอบเพื่อให้ได้จำนวนรอบฝึกซ้ำที่มากจนสามารถสรุปผลลัพธ์ได้อย่างมีนัยสำคัญทางสถิติ (p-value) นอกจากนี้เนื่องด้วยลักษณะเฉพาะของการฝึกอบรมแบบจำลองการเรียนรู้เชิงลึกที่ใช้ทรัพยากรการคำนวณมาก ทำให้การฝึกอบรมแบบจำลองหนึ่งครั้งสามารถกินเวลาฝึกอบรมนานหลายวัน หลายสัปดาห์ หรือหลายเดือนขึ้นกับทรัพยากรการคำนวณที่มี ความยากของปัญหา และปริมาณของข้อมูลฝึกอบรมอีกด้วย

ผลลัพธ์เมื่อสิ้นสุดการฝึกอบรมจำนวน 4,000 epochs คือ YOLOv3 ได้ Average Loss เท่ากับ 0.0947 ใช้เวลาในการฝึกอบรมเฉลี่ย 5 ชั่วโมง ส่วน YOLOv4 ได้ผลของ Average Loss เท่ากับ 0.7250 ใช้เวลาในการฝึกอบรมเฉลี่ย 14 ชั่วโมง อย่างไรก็ตามในการพิจารณาเลือกระหว่าง YOLOv3 และ YOLOv4 นั้นผู้วิจัยไม่ได้นำค่า Average Loss จากทั้งสองโมเดลดังกล่าวมาเป็นตัวชี้วัดในการเลือก เนื่องจาก Average Loss นั้นเป็นค่าที่ถูกเลือกมาให้เหมาะสมกับอัลกอริทึม Gradient Descent ที่ใช้ในการเรียนรู้เชิงลึก แต่ในกรณีของโมเดลตรวจจับวัตถุนั้นค่า Average Loss เป็นการบวกรวมกันของค่าความผิดพลาดหลายส่วนซึ่งไม่เหมาะสมจะนำมาใช้สะท้อนประสิทธิภาพและความคุ้มค่าในการใช้งานของโมเดลประเภทนี้ นอกจากนี้สถาปัตยกรรมและการคำนวณภายในของ YOLOv3 และ YOLOv4 ก็มีความแตกต่างกัน จึงไม่สามารถนำค่า Average loss จากทั้งสองโมเดลมาเปรียบเทียบข้ามกันได้

สำหรับตัวชี้วัดประสิทธิภาพในโมเดลการเรียนรู้เชิงลึกประเภทตรวจจับวัตถุนั้น สามารถสรุปจากงานวิจัยในอดีตได้ดังตารางที่ 5 โดยในงานวิจัยชิ้นนี้ผู้วิจัยเลือกใช้ตัวชี้วัดประสิทธิภาพจำนวน 5 ได้แก่ Average Precision (AP) สำหรับวัดความแม่นยำในการตรวจจับวัตถุ, F1-score สำหรับวัดความสามารถของโมเดล โดยเฉลี่ยค่า Precision และ Recall, mAP สำหรับวัดค่าเฉลี่ยของ Average Precision ของวัตถุทั้งหมด, Frame Per Seconds (FPS) สำหรับวัดประสิทธิภาพความเร็ว (Xiaowei et al., 2021) และAverage IoU สำหรับวัดประสิทธิภาพของโมเดล โดยหาจำนวนเปอร์เซ็นต์ที่ทับซ้อนกันระหว่างผลเฉลย และผลจากการทำนาย (Rattanachot, 2019) ทั้งนี้ตัวชี้วัดทั้งหมดที่กล่าวมายิ่งมีค่ามาก ยิ่งสะท้อนถึงประสิทธิภาพการทำงานที่ดีของโมเดล

ประเภทตรวจจับวัตถุ ในส่วนของค่า Precision และ Recall นั้นผู้วิจัยไม่ได้นำมาใช้เพราะการใช้ F1-Score เป็นค่าที่เฉลี่ย Precision และ Recall มาอยู่แล้ว นอกจากนี้งานวิจัยนี้ไม่มีการใช้ค่า False Positive Rate และ False Negative Rate ในการประเมินโมเดล เนื่องจากสำหรับโมเดลตรวจจับวัตถุนั้นกรอบภาพ (Bounding Box) ที่เป็นไปได้ทั้งหมดที่ไม่ใช่ False Negative (FN), False Positive (FP) และ True Positive (TP) ก็คือ True Negative (TN) ทำให้จำนวนของ TN มีมากและละเอียดเกินไป ค่า TN จึงไม่ถูกนำมาใช้วัดประสิทธิภาพ ส่งผลให้ไม่มีการนำค่า False Positive Rate และ False Negative Rate ซึ่งมีสูตรการคำนวณขึ้นอยู่กับ TN มาใช้ในการเปรียบเทียบโมเดลตรวจจับวัตถุในที่นี้

ตารางที่ 5 ตัวชี้วัดประสิทธิภาพโมเดลประเภทตรวจจับวัตถุ (Object Detection) จากงานที่เกี่ยวข้อง

งานวิจัย	ตัวชี้วัดประสิทธิภาพโมเดล Object Detection						
	Precision	**Recall**	**F1-Score**	**mAP**	**AP**	**Average IoU**	**FPS**
Small Object Detection in Traffic Scenes Based on YOLO-MXANet (Xiaowei et al., 2021)	✓	✓	✓	✓			
Fire-YOLO: A Small Target Object Detection Method for Fire Inspection (Lei et al., 2022)	✓	✓	✓	✓			
Comparing YOLOv3 , YOLOv4 and YOLOv5 for Autonomous Landing Spot Detection in Faulty UAVs (Upesh & Hossein, 2022)	✓	✓	✓	✓			
You Only Look Once (YOLOv3): Object Detection and Recognition for Indoor Environment (Hassan et al., 2021)	✓	✓	✓	✓	✓	✓	
YOLO-GD: A Deep Learning-Based Object Detection Algorithm for Empty-Dish Recycling Robots (Xuebin et al., 2022)	✓	✓	✓		✓		✓

ตารางที่ 6 แสดงการเปรียบเทียบประสิทธิภาพของ YOLOv3 และ YOLOv4 บนตัวชี้วัดประสิทธิภาพ 5 ตัวที่กล่าวสรุปไปข้างต้น พิจารณาจากค่าในตารางจะเห็นว่าเมื่อโมเดลถูกนำมาประเมินกับข้อมูลชุดทดสอบโดยภาพรวม YOLOv4 ให้ผลลัพธ์การตรวจจับอาวุธในภาพการ์ตูนได้ดีกว่า YOLOv3

ตารางที่ 6 เปรียบเทียบประสิทธิภาพของ YOLOv3 และ YOLOv4 ในการตรวจจับอาวุธบนข้อมูลชุดทดสอบ ตัวเลขที่เป็นตัวหนาและขีดเส้นใต้คือผลลัพธ์ที่ดีที่สุดของค่า Mean และ SD ในแถวหนึ่ง ๆ

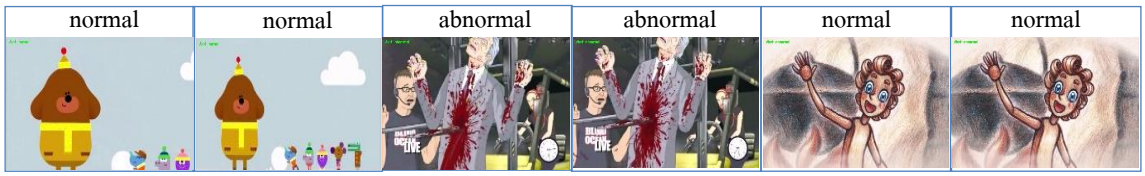
Performance on Testing Dataset	YOLOv3		YOLOv4	
	Mean	**S.D.**	**Mean**	**S.D.**
Average Precision (AP): Gun	91.88%	0.009	**95.79%**	**0.004**
Average Precision (AP): Knife	52.97	0.109	**59.39%**	**0.031**
F1-score	0.80	0.05	**0.84**	**0.005**
Average IoU	65.71%	0.039	**68.99%**	**0.024**
mAP@0.50	72.43%	0.059	**77.59%**	**0.014**
Average FPS (Tesla K80)	**30.73**	1.418	18.63	**0.777**

3. การประเมินผลกับวิดีโอการ์ตูน

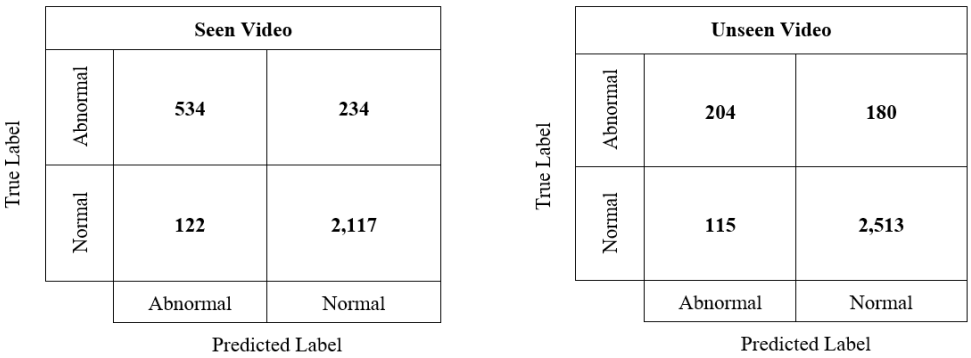
3.1 ประเมินผลการจำแนกภาพ

ในหัวข้อนี้ผู้วิจัยนำเอาโมเดลจำแนกภาพที่ดีที่สุดในงานนี้คือ VGG16 มาทำการทดสอบประสิทธิภาพเพิ่มอีก โดยเป็นการทดสอบเพื่อเปรียบเทียบว่าประสิทธิภาพในการจำแนกภาพของ VGG16 ตัวนี้จะเปลี่ยนแปลงไปหรือไม่อย่างไรเมื่อถูกนำไปใช้กับภาพในวิดีโอการ์ตูนที่โมเดลไม่เคยเห็นมาก่อน (Unseen) ซึ่งภาพเหล่านี้หมายถึงภาพที่ไม่มีปรากฏอยู่ทั้งในข้อมูลฝึกอบรม (Train Set) และข้อมูลทดสอบ (Test Set) นั่นเอง สำหรับการเตรียมข้อมูลเพื่อการทดลองนี้ในส่วนของวิดีโอการ์ตูนที่เคยเห็นมาก่อน (Seen) ผู้วิจัยนำเอาวิดีโอการ์ตูนความยาว 4 นาที 59 วินาทีมาทำการสกัดได้เฟรมภาพสำหรับใช้ทดลองจำนวน 3,007 ภาพ และในส่วนของวิดีโอการ์ตูนที่ไม่เคยเห็นมาก่อน (Unseen) ผู้วิจัยนำเอาวิดีโอการ์ตูนความยาว 5 นาทีมาสกัดได้เฟรมภาพสำหรับทดลองทั้งหมด 3,012 ภาพ เพื่อหาความถูกต้อง (Accuracy) ในการจำแนกภาพโดย VGG16 ทำการแยกภาพ ทำเป็น Confusion Matrix

ตัวอย่างผลการจำแนกภาพโดย VGG16 ของการทดลองนี้แสดงในรูปที่ 10 และค่า Confusion Matrix ของการจำแนกภาพเปรียบเทียบระหว่างวิดีโอการ์ตูนที่เคยเห็นและไม่เคยเห็นแสดงในรูปที่ 11 เมื่อคำนวณจากรูปที่ 11 จะพบว่าค่า Accuracy ของการจำแนกภาพโดย VGG16 นั้นมีค่าเป็น 88.16% และ 90.20% สำหรับวิดีโอที่เคยเห็นและไม่เคยเห็นตามลำดับ กล่าวคือในการทดลองนี้ VGG16 สามารถจำแนกภาพบนวิดีโอการ์ตูนที่ไม่เคยเห็นได้ดีกว่าซึ่งถือเป็นพฤติกรรมที่ผิดวิสัยของระบบการเรียนรู้ของเครื่องจักร (Machine Learning) ใด ๆ สาเหตุของพฤติกรรมที่ไม่ปกตินี้ ผู้วิจัยคาดว่าเกิดจากสัดส่วนจำนวนของภาพที่เหมาะสม (Normal) และไม่เหมาะสม (Abnormal) ที่แตกต่างกันในวิดีโอทั้งสอง โดยในวิดีโอที่เคยเห็นนั้นมีภาพเหมาะสมและไม่เหมาะสมอยู่ 2,239 ภาพ (74.46%) และ 768 ภาพ (25.54%) ตามลำดับ ในขณะที่วิดีโอที่ไม่เคยเห็นนั้นมีภาพเหมาะสมและไม่เหมาะสมอยู่ 2,628 ภาพ (87.25%) และ 384 ภาพ (12.75%) ตามลำดับ กล่าวคือ เนื่องจาก VGG16 สามารถจำแนกภาพเหมาะสม (Normal) ได้ดีกว่า ทำให้ประสิทธิภาพ Accuracy โดยรวมบนวิดีโอที่ไม่เคยเห็นซึ่งมีภาพเหมาะสมอยู่ถึง 87.25% ให้ผลลัพธ์ที่ดีกว่าวิดีโอที่เคยเห็นซึ่งมีภาพเหมาะสมอยู่เพียง 74.46%



รูปที่ 10 ตัวอย่างผลลัพธ์สำหรับการประเมินผลการจำแนกภาพในวิดีโอการ์ตูน (เครดิตภาพการ์ตูนระบุในหัวข้อ Data Collection)

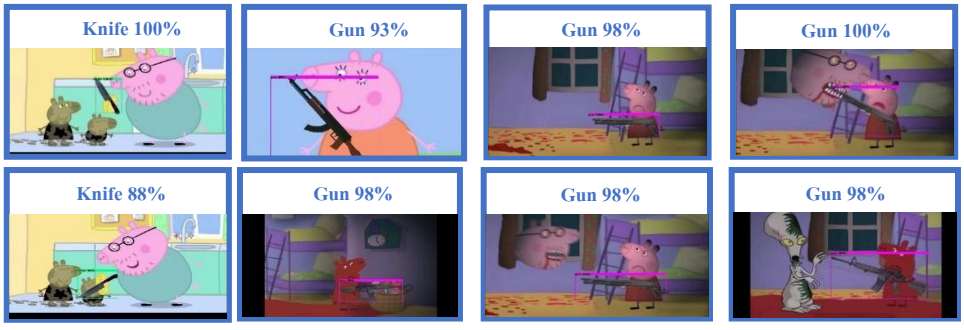


รูปที่ 11 Confusion Matrix ผลการจำแนกภาพโดย VGG16 บนวิดีโอการ์ตูนที่เคยเห็น (ซ้าย) และไม่เคยเห็น (ขวา)

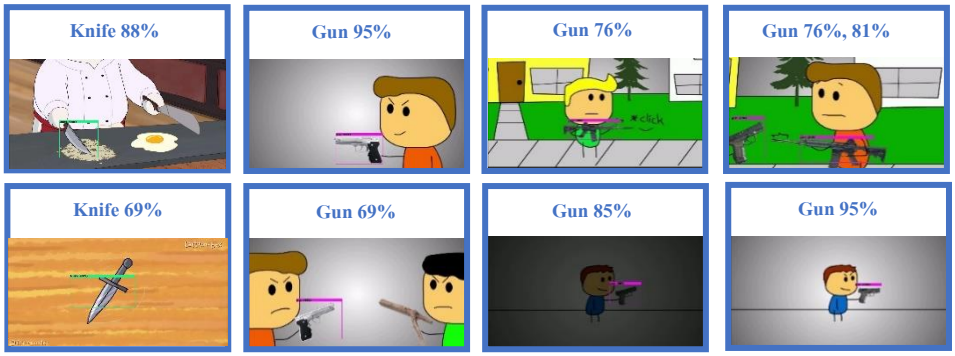
3.2 ประเมินผลการตรวจจับอาวุธ

ในส่วนนี้ผู้วิจัยนำโมเดล YOLOv4 ที่ถูกฝึกอบรมไว้แล้วจากหัวข้อก่อนหน้ามาทำการทดสอบเปรียบเทียบประสิทธิภาพบนวิดีโอการ์ตูนที่เคยเห็น (วิดีโอจาก Train Set) และไม่เคยเห็น (วิดีโอการ์ตูนที่ไม่อยู่ใน Train Set) ซึ่งมีความยาว 5 นาทีเท่ากัน ผลการทดลองบนวิดีโอที่เคยเห็น (ตัวอย่างในรูปที่ 12) พบว่าโมเดล YOLOv4 ใช้เวลาในการประมวลผลวิดีโอทั้งหมดนาน 19 นาที (1,140 วินาที) มีความเร็ว 8.0 fps (frame per second) และมีค่า Average Precision (AP) ของปืนและมีดเท่ากับ 91.44% และ 87.39% ตามลำดับ ส่วนค่า Mean Average Precision (mAP) โดยรวมนั้นอยู่ที่ 89.41% ในส่วนของผลบนวิดีโอการ์ตูนที่ไม่เคยเห็นนั้น (ตัวอย่างแสดงในรูปที่ 13) โมเดลใช้เวลาประมวลผลวิดีโอนาน 21 นาที (1,260 วินาที) มีความเร็ว 8.0 fps และมีค่า AP ของปืนและมีดคือ 68.8% และ 59.18% ตามลำดับ ในส่วนของค่า mAP นั้นอยู่ที่ 63.99%

กล่าวโดยสรุปคือความสามารถในการตรวจจับอาวุธของ YOLOv4 นั้นแย่ลงเมื่อถูกนำไปใช้กับวิดีโอการ์ตูนที่ไม่เคยเห็นมาก่อน วิธีการแก้ปัญหาแบบตรงไปตรงมาคือการเพิ่มจำนวนข้อมูลฝึกสอนให้ครอบคลุมความเป็นไปได้ของภาพการ์ตูนต่าง ๆ มากขึ้น



รูปที่ 12 การประเมินผลการตรวจจับอาวุธจากวิดีโอที่เคยเห็น (เครดิตภาพการ์ตูนระบุในหัวข้อ Data Collection)



รูปที่ 13 การประเมินผลการตรวจจับอาวุธจากวิดีโอที่ไม่เคยเห็น (เครดิตภาพการ์ตูนจาก Brewstew - Airsoft Guns [Video file], https://www.youtube.com/watch?v=ViNnMcOqDCA&t=60s)

วิจารณ์และสรุปผลการทดลอง

ในส่วนของการจำแนกภาพ ผู้วิจัยได้ทำการทดลองตั้งค่าโมเดลหลากหลายรูปแบบ โดยมีการนำเทคนิค Transfer Learning และ Finetuning ของการเรียนรู้เชิงลึกมาใช้กับโมเดล ResNet50, VGG16 และ VGG19 ซึ่งผลการทดลองพบว่าโมเดลที่ให้ผลลัพธ์ดีที่สุดกับชุดข้อมูลภาพการ์ตูนที่เตรียมไว้คือ VGG16 โดยเมื่อนำมาทดสอบซ้ำกับวิดีโอการ์ตูนที่เคยเห็นและไม่เคยเห็นก็พบว่าได้ความสามารถในการจำแนกภาพอยู่ที่ Accuracy 88.16% และ 90.20% ตามลำดับ ซึ่งผลลัพธ์ของวิดีโอไม่เคยเห็นที่ดีกว่านั้นเป็นผลพวงมาจากการที่ข้อมูลภาพในวิดีโอดังกว่ามีจำนวนของข้อมูลภาพเหมาะสม (Normal) ที่โมเดลถนัดในการจำแนกมากกว่านั่นเอง หากต้องการที่จะปรับการทดลองให้ได้ผลลัพธ์ที่แม่นยำมากขึ้นอีก ผู้วิจัยมีความเห็นว่าควรต้องรวบรวมภาพสำหรับใช้ในการทดลองให้มากขึ้น รวมถึงอาจเพิ่มเทคนิคการทำ Data Augmentation ให้มากและหลากหลายยิ่งขึ้นอีก ทั้งนี้อีกเหตุผลหนึ่งที่ผู้วิจัยคาดว่าเป็นสาเหตุที่ทำให้ผลลัพธ์การจำแนกภาพของบางโมเดลออกมาไม่ค่อยดีนั้น เป็นเพราะโมเดล ResNet50, VGG16 และ VGG19 ที่ผู้วิจัยหยิบมาใช้ต่างก็ถูกฝึกสอนมาก่อนแล้ว (pre-trained) ด้วยชุดข้อมูล ImageNet ซึ่งข้อมูลฝึกสอนกว่าหนึ่งล้านรูปใน ImageNet นั้นเป็นภาพถ่ายในชีวิตประจำวันที่ถูกโพสต์ลงบนโซเชียลมีเดีย ไม่ใช่ภาพลายเส้นอย่างภาพการ์ตูนที่ใช้ในงานวิจัยนี้ ด้วยลักษณะของภาพถ่ายและภาพการ์ตูนที่แตกต่างกันอย่างมีนัยสำคัญนี้เอง จึงอาจเป็นเหตุให้การ Transfer Learning และ Finetuning ให้ผลในบางโมเดลไม่ดีเท่าที่ควร

ในส่วนของการตรวจจับอาวุธในวิดีโอการ์ตูนด้วยโมเดล YOLOv3 และ YOLOv4 นั้น ผู้วิจัยพบว่าผลลัพธ์ค่า Average Precision ของมีดในทั้งสองโมเดลมีค่าต่ำ คาดว่าเป็นเหตุมาจากจำนวนภาพมีดที่ผู้วิจัยเก็บรวบรวมมาได้น้อยเกินไป หากสามารถเพิ่มภาพมีดในข้อมูลฝึกอบรมให้มากขึ้น น่าจะได้ค่า Average Precision ของมีดที่สูงขึ้นเทียบเท่ากับปืน และในการทดลองใช้ YOLOv4 ซึ่งเป็นตัวที่ให้ผลลัพธ์ดีกว่ากับวิดีโอที่เคยเห็นเปรียบเทียบกับวิดีโอที่ไม่เคยเห็นนั้น พบว่า YOLOv4 สามารถตรวจจับอาวุธบนวิดีโอที่เคยเห็นได้ค่อนข้างแม่นยำที่ Mean Average Precision เท่ากับ 89.41% แต่เมื่อมาทดสอบกับวิดีโอที่ไม่เคยเห็นค่าก็ตกไปเหลือ 69.33% โดยเมื่อโมเดลเจอภาพการ์ตูนที่มีรูปทรงของอาวุธไม่เหมือนกับปืนจริงหรือปืนที่เป็นภาพวาด โมเดลจะไม่สามารถตรวจจับปืนการ์ตูนนั้นได้ รวมถึงหากปืนหรือมีดอยู่ไกลมากหรือมีขนาดเล็กมากก็ทำให้โมเดลไม่สามารถตรวจจับได้เช่นกัน การแก้ปัญหาทั้งสองนี้นอกจากจะเพิ่มการฝึกอบรมด้วยภาพปืนและมีดแบบต่าง ๆ ให้ครอบคลุมมากขึ้นแล้ว ยังรวมถึงการพิจารณาและทดลองใช้สถาปัตยกรรมโมเดลแบบอื่น ๆ ที่มีความสามารถตรวจจับวัตถุขนาดเล็กในภาพได้ดี

ภายภาคหน้าของการศึกษาเพื่อต่อยอด ผู้วิจัยจะเพิ่มข้อมูลโดยการนำภาพลายเส้นจากเกมคอมพิวเตอร์มาเสริมให้มีข้อมูลในการฝึกสอนมากขึ้น และนำงานวิจัยนี้ไปทดสอบวิเคราะห์ความเหมาะสมของสื่อวิดีโอการ์ตูนสำหรับเด็กในประเทศไทย

เอกสารอ้างอิง

Burugupalli, M. (2020*). Image classification using transfer learning and convolution neural networks* [Unpublished master's thesis]. North Dakota State University.

Dyer, T. (2018). The Effects of Social Media on Children. *Dalhousie Journal of Interdisciplinary Management, 2018*(14), 1-16.

Electronic Transactions Development Agency. (2020). *Thailand internet user behavior 2020*. https://www.etda.or.th/th/newsevents/pr-news/ETDA-released-IUB-2020.aspx, Cited 19 July 2021

Hassan, S., Hassan, J., & Salma, H. (2021). You Only Look Once (YOLOv3): Object Detection and recognition for indoor environment. *Multicultural Education*, *7*(6), 171–181. https:\\doi\10.5281/zenodo.4906284

He, X., Cheng, R., Zheng, Z., & Wang, Z. (2021). Small object detection in traffic scenes based on YOLO-MXANet. *Sensors*, *21*(21), 7422. https://doi.org/10.3390/s21217422

Jain, N., Gupta, V., Shubham, S., Madan, A., Chaudhary, A., & Santosh, K. C. (2021). Understanding cartoon emotion using integrated deep neural network on large dataset. *Neural Computing and Applications*. https://doi.org/10.1007/s00521-021-06003-9

Khan, N. F., Hussain, A., Khan, A., & Din, I. U. (2019). *Categorized Violent Contents detection in cartoon movies using Deep Learning Model: Mobile Net*. *In The 5th International Conference on Next Generation Computing 2019, 20 - 21 December 2019, University of Petroleum & Energy Studies*, India.

Lamas, A. C., Pérez, F., & Gondim, J. (2021). *OD-WeaponDetection*. https://github.com/ari-dasci/OD-WeaponDetection

Lamoonmorn, Y., & Sucontphunt, T. (2019). *Full helmet and gun detection for robbery alarming from CCTV data in real time*. Bangkok: National Institute of Development Administration. (in Thai)

Narejo, S., Pandey, B., Esenarro Vargas, D., Rodriguez, C., & Anjum, M. R. (2021). Weapon detection using YOLO V3 for smart surveillance system. *Mathematical Problems in Engineering*, *2021*, 1-9. https://doi.org/10.1155/2021/9975700

Nepal, U., & Eslamiat, H. (2022). Comparing YOLOv3, YOLOv4 and YOLOv5 for autonomous landing spot detection in faulty UAVs. *Sensors*, *22*(2), 464. https://doi.org/10.3390/s22020464

Ratchatathanarat, T., & Wongcharoen, N. (2021). Influence of media on violent behavior of child and youth in Chiangmai province. *Journal of Health Science, 30*(2), 199-210. (in Thai)

Rattanachot, P. (2019). *Automated plant disease detection using drones and deep leaning* [Unpublished master's thesis]. Dhurakij Pundit University. (in Thai)

Royal Thai Government Gazette, Vol. 130, Pt 147 Gnor, 29th October B.E. 2556 (A.D.2013), 21.

Tahir, R., Ahmed, F., Saeed, H., Ali, S., Zaffar, F., & Wilson, C. (2019). *Bringing the Kid back into YouTube Kids: Detecting Inappropriate Content on Video Streaming Platforms*. *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, (pp. 464-469). https:\\doi\10.1145/3341161.3342913

Yue, X., Li, H., Shimizu, M., Kawamura, S., & Meng, L. (2022). YOLO-GD: a deep learning-based object detection algorithm for empty-dish recycling robots. *Machines*, *10*(5), 294. https://doi.org/10.3390/machines10050294

Zhao, L., Zhi, L., Zhao, C., & Zheng, W. (2022). Fire-YOLO: a small target object detection method for fire inspection. *Sustainability*, *14*(9), 4930. https://doi.org/10.3390/su14094930