บทความวิจัย

โปรแกรมวิเคราะห์อารมณ์บนใบหน้าจากวิดีโอ

Facial emotion analysis program from video

ชัชวาลย์ เดินรัมย์, จิรายุ ช่างปรุง, ผศ.ดร.รพีพร ช่ำชอง
สาขาวิทยาการคอมพิวเตอร์ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม
62011212001@msu.ac.th, 62011212034@msu.ac.th, rapeeporn@msu.ac.th

บทคัดย่อ

เทคโนโลยีการตรวจจับวัตถุที่สามารถค้นหาสิ่งต่างๆ ที่อยู่ในรูปภาพหรือวิดีโอ โดยสามารถสอนโมเดลการตรวจจับฝห้จดจำแค่บางสิ่งเพื่อใช้ในบางงานที่เจาะจงได้ เทคโนโลยีการตรวจจับวัตถุในปัจจุบันเริ่มมีให้เห็นทั่วๆ ไปแล้ว เช่นกล้องวงจรปิด มือถือ รถยนต์ เป็นต้น เมื่อเทคโนโลยีการตรวจจับวัตถุสามารถนำไปใช้ได้ในหลายๆ งาน ดังนั้นในงานวิจัยนี้จึงใช้โมเดลการเรียนรู้เชิงลึก (Deep learning) ได้เรียนรู้ลักษณะใบหน้าในแต่ละอารมณ์โดยใช้โมเดลที่ผ่านการเรียนรู้มาแล้วที่มีพื้นฐานมาจาก CNN (Convolutional Neural Network) ที่มีชื่อว่า VGG-16 และใช้หลักการตรวจจับวัตถุแบบ Faster R-CNN ที่มีการสร้างโมเดลแยกออกมาอีกหนึ่งขั้นตอนเพื่อทำงานในการคัดเลือกภาพก่อนจะส่งไปทำนายผลจริงชื่อว่า RPN (Region Proposal Network) โดยการเรียนรู้และทดสอบจากภาพใบหน้าคนจริงๆ ผลการวัดประสิทธิภาพได้ให้ผลที่น่าพึงพอใจและนำโมเดลปรับใช้ในงานหลายๆ ด้าน

คำสำคัญ : การตรวจจับวัตถุ, Faster R-CNN, Region Proposal Network, Convolutional Neural Network)

1.บทนำ

ในปัจจุบัน การวิเคราะห์อารมณ์บนใบหน้าจากวิดีโอถูกใช้งานอย่างแพร่หลาย การวิเคราะห์อารมณ์บนใบหน้าสามารถนำไปใช้ได้กับงานหลากหลายแขนง เช่น งานด้านรู้จำใบหน้า งานด้านการสร้างออกแบบตัวละครเพื่อไปวิเคราะห์อารมณ์สีหน้าท่าทางของตัว งานด้านการบริการของพนักงานกับลูกค้าของบริษัทต่างๆ และธนาคารก็เช่นกัน การวิเคราะห์อารมณ์บนใบหน้าจากวิดีโอถูกนำไปประยุกต์ใช้ในกับงานด้านต่างๆ เพื่อต่อยอดขึ้นไป

ในสถานการณ์โลกปัจจุบัน โควิด–19 (Covid-19) เข้ามามีผลกระทบกับการเรียนการสอน และการพูดคุยขอคำปรึกษากันระหว่างผู้สอนและผู้เรียน จึงทำให้ต้องติดต่อสื่อสารกันทางวิดีโอคอล (Video Call) ซึ่งบางครั้งก็ไม่ได้ดูหน้าจอทำให้ไม่สามารถรู้ความรู้สึกต่างๆ ของผู้เรียนหรือของผู้ขอคำปรึกษาจากอาจารย์ที่ปรึกษา บางครั้งผู้เรียนก็อาจจะไม่สนุกกับการเรียนหรือเรียนไม่เข้าใจ จะช่วยให้ผู้สอนนำไปปรับปรุงการสอนให้ดียิ่งขึ้น

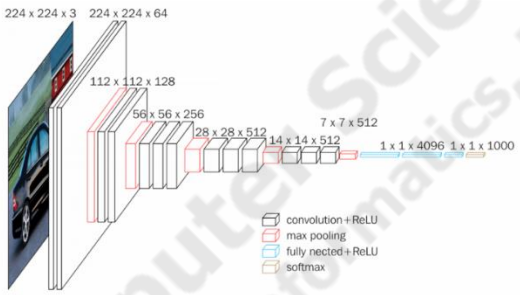
ดังนั้นทางผู้จัดทำจึงออกแบบโปรแกรมวิเคราะห์อารมณ์บนใบหน้าจากวิดีโอ

เพื่อช่วยให้ผู้สอนสามารถนำผลลัพธ์ที่ได้ไปปรับปรุงการสอนให้ดียิ่งขึ้น หรือการให้คำปรึกษาของที่ปรึกษากับผู้ขอคำปรึกษา และสามารถนำผลการวิเคราะห์อารมณ์บนใบหน้าไปประยุกต์ใช้งานกับด้านอื่นๆ ต่อไปได้

2.ทฤษฎีและงานที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

Convolutional Neural Network (CNN) [5] หรือ โครงข่ายประสาทแบบคอนโวลูชันเป็นจำลองการมองเห็นของมนุษย์ที่มองเป็นส่วนย่อยๆ และนำกลุ่มของส่วนย่อย ๆ มาผสานกัน เพื่อดูว่าสิ่งที่เห็นอยู่คืออะไร โดยใช้ค่าพิเซลที่ได้จากข้อมูลอินพุต มีทั้งหมด 3 สี ได้แก่ สีแดง, น้ำเงิน, และเขียว สามารถใช้เลข 0 ถึง 255 เพื่อแทนค่าความเข้มของสี



ภาพประกอบที่ 1 Convolutional Neural Network VGG-16

รับภาพ Input เข้ามาเป็น array ขนาด 224*224*n โดยที่ n คือจำนวนโหนด depth จากนั้น

ทำ max pooling เพื่อหาค่าที่มากที่สุดของจุดภาพด้วยตัวกรอง(filter) ขนาด 3*3 คือการลดขนาดของจุดภาพโดยที่ให้สูญเสียรายละเอียดของภาพน้อยที่สุด แล้วก็ประมวลผลแบบนี้ไปเรื่อย ๆ ตามจำนวนของโหนด Hidden layer ไปจนถึงชั้น Fully Connected (FC) ส่งค่าคำนวณจากโหนดหนึ่ง ไป

อีกโหนดหนึ่ง เชื่อมต่อ Hidden layer ต่างๆ เข้าด้วยกัน แล้วทำการประมวลผล output ออกมาตาม

Class จากนั้น เมื่อถึงชั้น Soft max ก็จะแปลงค่าของ output ออกมาให้อยู่ในรูปแบบความน่าจะเป็น

โครงสร้างของ Convolutional Neural Network ประกอบได้ดังนี้

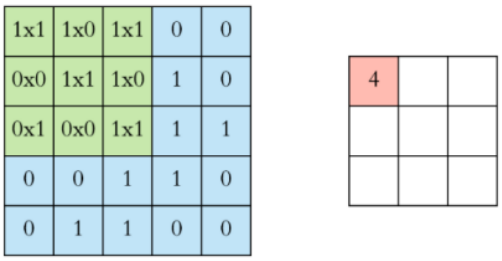
(1) Convolutional

เป็น Layer หลักของ CNN ทำหน้าที่รับ Input เข้ามา แปลงภาพให้เป็นพิกเซล ที่กำหนดให้

เป็น 0 ถึง 255 จากนั้นจะใช้การดำเนินการทางคณิตศาสตร์เพื่อหาคุณสมบัติที่สำคัญจากรูปภาพเตอร์

การคำนวณจะเริ่มจากการกำหนดค่าใน ตัวกรอง (filter) หรือ เคอร์เนล (kernel) ที่ช่วยดึงคุณลักษณะ

ที่ใช้ในการรู้จำวัตถุออกมา หรือที่เรียกว่า Feature Map



ภาพประกอบที่ 2 Feature Map

การทำงานของ CNN จะทำการ Sliding Windows (Filter) เพื่อค้นหาองค์ประกอบของภาพเช่น สี หรือรูปร่าง

$$output\ of\ size = \frac{N-F+2P}{S} + 1$$

โดยที่ N คือ ขนาดของภาพ

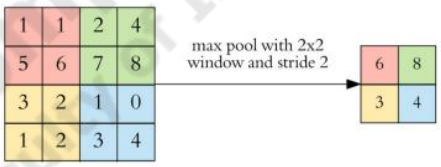
F คือ ขนาดของ Filter

P คือ จำนวนของ Padding

S คือ จำนวนของ Stride (จำนวนของการขยับ Filter)

(2) Pooling Layer

เป็นชั้นที่เชื่อมจาก Convolutional Layer โดยมีเป้าหมายคือทำให้ขนาดของ Feature Map ลดลงด้วยการหาค่าเฉลี่ย (Average Pooling) หรือหาค่าที่สูงที่สุด (Max Pooling) และ
จะเลื่อนตัวกรองไปตาม Stride ที่กำหนดไว้ โดยขนาดตัวกรองของการทำ Max Pooling นิยมเรียกกัน
ว่า Pool Size



ภาพประกอบที่ 3 Pooling Layer

(3) Fully Connected layer

โดยขั้นตอนการหาค่าแต่ละโหนด ในขั้นตอน Fully Connected layer

$$H_i = \sum_{i=o}^{n-1}(x_i \cdot W_i)$$

โดยที่ H_i คือ ผลลัพธ์ Hidden Layer โหนดที่ i

n คือ จำนวน Input ของโหนดก่อนหน้า

x_i คือ ข้อมูลของโหนด Input

W_i คือ ค่าน้ำหนัก

และเมื่อได้ผลลัพธ์นำข้อมูลเข้าฟังก์ชันที่รับผลรวมการประมวลผลทั้งหมด Sigmoid Function

$$F(A) = \frac{1}{1+e^{-A}}$$

โดยที่ F คือ ผลลัพธ์ Sigmoid Function มีค่าระหว่าง 0 ถึง 1

A คือ ผลลัพธ์ของ Hidden Layer

2.1.1 VGG16

เป็นหนึ่งในโมเดลที่มีพื้นฐานมาจากโครงสร้างประสาทเทียมแบบคอนโวลูชั่นที่มีการแข่งขันด้วยชุดข้อมูล ImageNet และติดอันดับโมเดลห้าอันดับที่ดีที่สุดโดยการทำงานมีโครงสร้างดังนี้

ตารางที่ 1 โครงสร้าง VGG-16

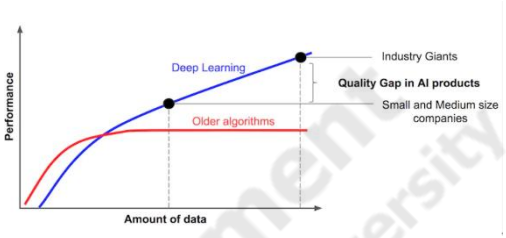
Input	Layer
1	2 X Convolution
	Max Pooling

ตารางที่ 1 โครงสร้าง VGG-16(ต่อ)

Input	Layer
3	2 X Convolution
	Max Pooling
5	2 X Convolution
	Max Pooling
7	3 X Convolution
	Max Pooling
10	3 X Convolution
	Max Pooling
13	FC
14	FC
15	FC
Output	FC

2.1.1 การขยายข้อมูล

การขยายข้อมูลเป็นการเพิ่มจำนวนข้อมูลให้มีจำนวนมากขึ้นเพื่อให้เพียงพอต่อการนำไปใช้ในการทำงานกับรูปภาพการขยายข้อมูลหมายถึงการเพิ่มจำนวนของรูปภาพในฐานข้อมูล ในการทำงานกับข้อมูลในลักษณะปกติจะหมายถึงการเพิ่มแถวข้อมูล การขยายข้อมูลถูกนำเนื่องจากมนุษย์มีจำนวนข้อมูลที่จำกัด แต่ตามหลักการแล้วยิ่งมีข้อมูลมากขึ้นโมเดลของ Machine Learning ก็จะมีประสิทธ์ภาพดีขึ้น อย่างไรก็ตามการประมวลผลข้อมูลในทุกแบบย่อมเกี่ยวพันกับค่าใช้จ่าย ค่าใช้จ่ายนี้อาจจะเป็นในรูปแบบเงินตรา การลงแรงของมนุษย์ หรือพลังการประมวลผลของคอมพิวเตอร์ และแน่นอนว่าจะต้องเสียเวลาในการประมวลผล ดังนั้นเราจึงต้องการการขยายข้อมูลที่มีอยู่แล้วเพื่อเพิ่มจำนวนของข้อมูลที่จะป้อนให้โมเดล Machine Learning เพื่อให้โมเดลสามารถทำหน้าที่ได้อย่างมีประสิทธิภาพ



ภาพประกอบที่ 4 ความสัมพันธ์ระหว่างปริมาณของข้อมูล

การขยายข้อมูลสามารถทำได้หลายวิธี ในการทำงานกับรูปภาพจะใช้ การหมุนรูปภาพเดิม เปลี่ยนสภาพแสงในภาพ ครอบตัดภาพให้ลักษณะต่างออกไป ดังนั้นภาพหนึ่งภาพสามารถสร้างเป็น ข้อมูลภาพที่แตกต่างกันหลายๆ ภาพได้ตามเทคนิคที่ใช้ในการขยายข้อมูล ด้วยวิธีนี้เองเราสามารถลดปัญหาการ Overfit ของโมเดล Machine Learning กล่าวคือปัญหาที่ตัวแบบทำงานได้แม่นยำมากกับข้อมูลรูปภาพที่ใช้ในการฝึก แต่ทำงานได้ไม่แม่นยำในข้อมูลจริงซึ่งเป็นข้อมูลที่ตัวแบบไม่เคยเรียนรู้มาก่อน ในทางกลับกันถ้าใช้เทคนิคที่มีการ over sampling เช่น SMOTE จะทำให้มีโอกาสที่จะเกิด Overfit ซึ่งเป็นสิ่งที่ควรหลีกเลี่ยง

2.2 งานวิจัยที่เกี่ยวข้อง

งานวิจัยของ ธนพล พุ่มลำเจียก, 2016 [1] เรื่อง " FACIAL EXPRESSION RECOGNITION FROM VIDEO SEQUENCE USING LOCAL GABOR FILTERS AND PCA PLUS LDA " [1] เป็นการวิจัยที่ทำการรู้จำ

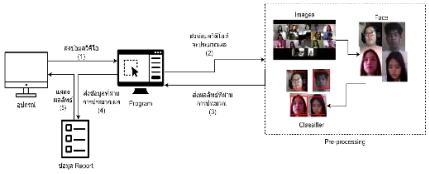
อารมณ์บนใบหน้าจากวีดีโอ โดยใช้ตัวกรองกาบอร์ วิธีการวิเคราะห์ องค์ประกอบหลักและการวิเคราะห์จำแนกประเภทเชิงเส้น โดยผลการทดลองเมื่อทาการเปรียบเทียบค่าความแม่นยำ ที่ได้จะพบว่าในฐานข้อมูล CK+ อัลกอริทึมที่ถูกพัฒนาขึ้นมีความแม่นยำ 94.58% ซึ่งมากกว่าการใช้การวิเคราะห์องค์ประกอบหลัก และการวิเคราะห์จำแนกประเภทเชิงเส้น อยู่ 10.16% และ 11.08% ตามลำดับ และใน ฐานข้อมูล JAFFE อัลกอริทึมที่ถูกพัฒนาขึ้นมีความแม่นยำ 97.5% ซึ่งมากกว่าการวิเคราะห์ องค์ประกอบหลักและการวิเคราะห์จำแนกประเภทเชิงเส้น อยู่ 7.63% และ 14.37%
ตามลำดับ และค่าความแม่นยำในการจำแนกระหว่างอารมณ์ปกติและอารมณ์โกรธ อัลกอริทึมที่
พัฒนาขึ้นมีความแม่นยำ 95% ซึ่งมากกว่า 20%

งานวิจัยของ จุฑามาศ มาบรรดิษ และคณะ, 2016 [2] เรื่อง “การรู้จำอารมณ์บนใบหน้า โดยใช้วิธีการวิเคราะห์องค์ประกอบหลัก และการวิเคราะห์จำแนกประเภทเชิงเส้น” เป็นงานวิจัยพัฒนาระบบรู้จำอารมณ์บนใบหน้าเพื่อต้องการให้ระบบดังกล่าวมีประสิทธิภาพและมีความถูกต้องมากยิ่งขึ้น อีกทั้งยังต้องการให้ระบบดังกล่าวมีความแพร่หลายมากขึ้นในประเทศไทย เนื่องจากกระบวนการในการวิเคราะห์อารมณ์นั้น ยังมีความซับซ้อนในการพัฒนา ผลการทดลองได้ทำการทดลองการรู้จำอารมณ์บนใบหน้าโดยใช้ PCA และ LDA ทดสอบกับฐานข้อมูลสามฐานข้อมูล คือ JAFFE, CK และ CPEKPS ปรากฏว่า PCA และ LDA ให้ค่าความถูกต้องที่ใกล้เคียงกันมาก โดย PCA ให้ผลที่ดีกว่า LDA เล็กน้อย และ L2-norm ให้ผลดีกว่า L1-norm

งานวิจัยของ Keyur Patel และ คณะ, 2020 [3] เรื่อง “Facial Sentiment Analysis using AI” เป็นงานวิจัยนำเสนอการสำรวจอย่างเป็นระบบโดยละเอียดเพื่อวิเคราะห์วิธีการที่ทันสมัยในปัจจุบันสำหรับการจดจำอารมณ์ใบหน้าในภาพนิ่งและพารามิเตอร์ต่าง ๆ ที่มีอิทธิพลต่อผลลัพธ์ของวิธีการเหล่านี้ เราได้พัฒนาระบบภาษีตามวิธีการที่แตกต่างกันที่ใช้สำหรับการตรวจจับใบหน้าการสกัดคุณลักษณะและการจำแนกอารมณ์ ผลการทดลองเราได้เปรียบเทียบวิธีการตรวจจับการสกัดและการจำแนกประเภทต่างๆและสรุปว่าวิธีการใดมีความโดดเด่นมากขึ้นในการบรรลุประสิทธิภาพที่ดีขึ้นในพลังการคำนวณที่มีอยู่ โดยการหารือเกี่ยวกับปัญหาและการวิจัยในปัจจุบันความท้าทายในอนาคตเราสรุปว่ายังมีการวิจัยที่จำเป็น

3.วิธีดำเนินงานวิจัย

3.1 ภาพรวมของระบบ



ภาพประกอบที่ 5 ภาพรวมระบบ

แผนผังการทำงานของระบบโดยที่มีผู้ใช้นำวิดีโอหรือรูปภาพส่งมาในโปรแกรม จากนั้นโปรแกรมจะนำวิดีโอหรือรูปภาพมาทำนายผลด้วย Deep Learning ทุกภาพ จากนั้นจะนำผลทำนายททุกๆ 10 นาทีออกมาใส่ในเอกสาร PDF ออกมา

3.2 ขั้นตอนการจัดเตรียมชุดข้อมูล

ภาพประกอบที่ 6 ตัวอย่างใบหน้าจากวิดีโอ

3.1.1 การเก็บรวบรวมข้อมูล

ข้อมูลที่ได้นำมาใช้เริ่มต้นเป็นไฟล์วิดีโอที่มาจาก YouTube วิดีโอยที่อัดเอง และวิดีโอจากอาจารย์ผู้สอนโดยมีคุณสมบัติดังนี้

1) ไฟล์วิดีโอนามสกุล mp4

2) ความคมชัด 1 ล้านพิกเซล หรือ 720P

3) มีค่าอัตราภาพต่อวินาทีหรือ FPS (Frame Per Second) เท่ากับ 30

หลังจากนั้นไปทำการแปลงข้อมูล เนื่องจากข้อมูลที่มีเป็นไฟล์วิดีโอจึงต้องทำการแปลงเป็น ไฟล์รูปภาพเพื่อใช้ในการนำไป Train ตัวแบบ โดยจากคุณสมบัติข้อมูลที่มีค่า FPS เท่ากับ 30 เพื่อทำการแปลงวิดีโอเป็นไฟล์ภาพนามสกุล jpg โดยกำหนดจัดเก็บ ภาพทุก 1 วินาทีของวิดีโอ โดยภาพที่จะนำไปใช้ฝึกสอนตัวแบบนำมาจากวิดีโอที่มาจาก YouTube วิดีโอที่อัดเอง และวิดีโอจากอาจารย์ผู้สอน จะทำการจัดเตรียม รูปภาพขนาดเต็ม (มุมกว้าง) ตามต้นฉบับของ วิดีโอโดยทั้งหมดได้จากการแคปภาพนิ่งออกมาที่ 1 FPS

3.1.2 การเตรียมข้อมูล

1) การทำความสะอาดข้อมูล

เมื่อทำการจัดเก็บรวบรวมข้อมูลรูปภาพที่ได้จากวิดีโอเรียบร้อย จะทำการทำความสะอาดข้อมูล (Data Cleansing) เนื่องจากรูปภาพที่ได้จากการแปลงผ่านวิดีโอมีข้อมูลที่ไม่จำเป็นต่อกระบวนการลำดับถัดไปจึงต้องทำการเอาออกจากชุดข้อมูล เช่น รูปภาพซ้ำกัน รูปภาพที่ไม่มีใบหน้าอยู่ ได้ผลลัพธ์ทั้งหมดจำนวนทั้งหมด 484 ภาพ ออกเป็นรูปภาพขนาดเต็ม (มุมกว้าง) ตามต้นฉบับของวิดีโอ

2) การแบ่งชุดข้อมูล (Splitting Data)

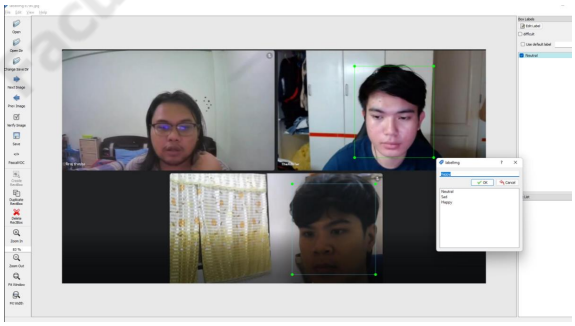
ทำการแบ่งสัดส่วนของข้อมูลเพื่อนำไปใช้ฝึกสอนตัวแบบ โดยจะแบ่งออกเป็นชุดข้อมูล

ฝึกสอน (Training) กับชุดข้อมูลทดสอบ (Testing) ซึ่งในงานวิจัยนี้จะทำการเลือกกลุ่มตัวอย่างอย่าง

ง่าย (Simple Random Sampling) โดยใช้คอมพิวเตอร์ในการสุ่ม ทำการแบ่งข้อมูลเป็น Training
และ Testing ในอัตราส่วน 80:20 ซึ่งคำนวณเป็นจำนวน 387 ภาพ และ 97 ภาพ ตามลำดับ

3) การกำกับประเภทของวัตถุบนรูปภาพ (Image Annotations)

ในขั้นตอนนี้จะทำการกำกับประเภทของวัตถุกับชุดข้อมูล Training เพื่อที่จะนำไปฝึกสอนตัว
แบบในการตรวจจับวัตถุ โดยจะทำการระบุขอบเขตของประเภทที่สนใจที่อยู่ภายในรูปภาพ เพื่อที่จะ
ทำให้ระบบค้นหาคุณลักษณะที่สำคัญในขอบเขตนั้น ๆ ซึ่งจะประกอบไปด้วยข้อมูลสำคัญสำหรับการ
ฝึกสอนตัวแบบดังต่อไปนี้ 1) ชื่อของประเภทวัตถุ 2) พิกัดแกน x และ y ของตำแหน่งที่ทำการ
กำกับไว้ (Bounding Box) ซึ่งจะถูกสร้างออกมาเป็นไฟล์ xml โดยในขั้นตอนนี้จำเป็นต้องลากกรอบ
ขอบเขตด้วยมือทีละภาพจนครบ 387 ภาพ ซึ่งเป็นขั้นตอนที่ใช้เวลานานในการดำเนินการ

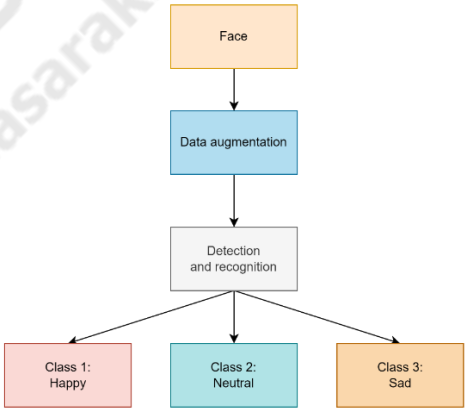


ภาพประกอบที่ 7 ตัวอย่างการกำกับขอบเขต

```
<annotation>
      <folder>samples</folder>
      <filename>abhi.jpg</filename>
      <path>E:/Documents Pro/Pro 1 CS/Face Recognition Code/images/samples/abhi.jpg</path
      <source>
            <database>Unknown</database>
      </source>
      <size>
            <width>297</width>
            <height>387</height>
            <depth>3</depth>
      </size>
      <segmented>0</segmented>
      <object>
            <name>Neutral</name>
            <pose>Unspecified</pose>
            <truncated>0</truncated>
            <difficult>0</difficult>
            <bndbox>
                  <xmin>61</xmin>
                  <ymin>8</ymin>
                  <xmax>231</xmax>
                  <ymax>240</ymax>
            </bndbox>
      </object>
</annotation>
```

ภาพประกอบที่ 8 ไฟล์ xml ที่ได้จากการทำ

3.2 ขั้นตอนการพัฒนาโมเดล



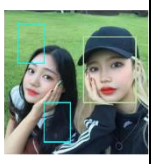
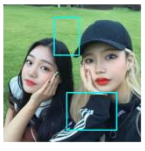
ภาพประกอบที่ 9 ขั้นตอนการพัฒนาของโมเดล

ในการพัฒนาโมเดลรู้จำอารมณ์ จะใช้การเรียนรู้แบบ Convolutional Neural Network VGG-16 โดยขั้นตอนการเรียนรู้โมเดลจะเป็นไปตามรูปภาพ ซึ่งจะทำการ Data Augmentation ซึ่งการเรียนรู้ของโมเดลมีขั้นตอนการทำงานดังนี้

1.Data Augmentation คือการสร้าง data ขึ้นมาใหม่ นำเอารูปเดิมมากลับซ้ายขวา

มา rotate มุมต่าง ๆ มาทำให้ภาพเบลอ ยิ่งมี data ที่หลากหลายเท่าไหร่ model ก็จะดีขึ้น ปัญหา overfit น้อยลง โดยจะทำอยู่ 8 ได้แก่ Add, Multiply, Dropout, GaussianBlur, Grayscale, GammaContrast, Fliplr, Flipud 2.Detection and recognition จะแบ่งออกเป็น 3 Class คือ 1) อารมณ์ดี (Happy) 2) อา-มรณ์ปกติ (Neutral) และ 3) อารมณ์เศร้า (Sad)

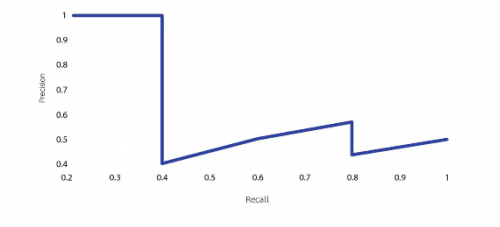
ตารางที่ 2 การคำนวณหาประสิทธ์ภาพ

ลำดับ	ตัวอย่างการ Predict	Precision $\frac{TP}{(TP+FP)}$	Recall $\frac{TP}{(TP+FN)}$
A		$\frac{1}{1+2}=0.33$	$\frac{1}{1+1}=0.5$
B		$\frac{2}{2+0}=1$	$\frac{2}{2+0}=1$
C		$\frac{0}{0+2}=0$	$\frac{0}{0+2}=0$

ตารางที่ 3 ตัวอย่างวัตถุหลายวัตถุ

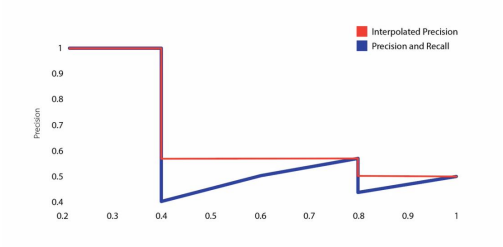
Confident	TP	FP	Cumulative TP	Cumulative FP	Precision	Recall
99%	1	0	1	0	1.0	0.2
98%	1	0	2	0	1.0	0.4
95%	0	1	2	1	0.67	0.4
87%	0	1	2	2	0.5	0.4
75%	0	1	2	3	0.4	0.4
60%	1	0	3	3	0.5	0.6
55%	1	0	4	3	0.57	0.8
45%	0	1	4	4	0.5	0.8
33%	0	1	4	5	0.44	0.8
30%	1	0	5	0	0.5	1.0

โดยที่ใน 1 ภาพจะมีหลายใบหน้าจะทำการเรียง Confident จากมากไปน้อย โดย Confident ที่ 99% จะนับ TP ทั้งหมด และ FP ทั้งหมด พร้อมกับหาค่า Precision และ Recall



ภาพประกอบที่ 10 กราฟจากตาราง

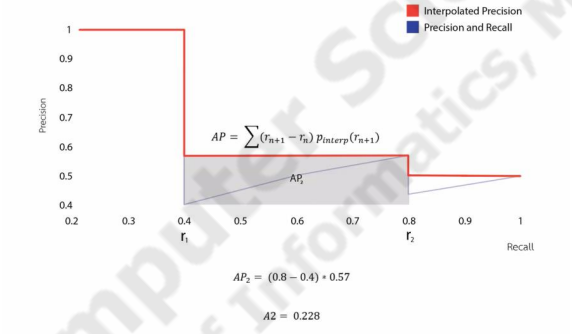
สามารถสร้างกราฟได้ทั้งต่อไปนี้ โดยเราต้องการหาพื้นที่ใต้กราฟ จะเป็นสัดส่วนความถูกต้องทั้งหมดที่ทำนายออกมาได้ทั้งหมด โดยต้องทำการ Interpolated Precision ให้คำนวณหาพื้นใต้กราฟได้ง่ายขึ้นดังภาพ



ภาพประกอบที่ 11 Interpolated Precision

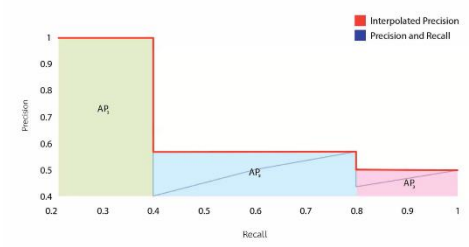
โดยทำการเปรียบเทียบค่า Precision ที่อยู่ทางด้านหน้าว่า มีค่า Precision ที่มากกว่าหรือไม่ ถ้าหากมีมากกว่า ก็จะทำการเติมค่า Precision ให้เท่ากับค่าที่อยู่ด้านหน้าเพื่อให้การคำนวณที่ง่ายขึ้น

โดยกราฟ Interpolated average precision ที่ได้จะมีลักษณะคล้ายคลึงกับกราฟเดิมแต่จะเติมค่า Precision ให้ทำการคำนวนได้ง่ายขึ้น



ภาพประกอบที่ 12 Confusion Matrix

จะทำการหาพื้นที่ใต้กราฟจากสมการที่กำหนดไว้ โดยที่ AP คือ พื้นที่ใต้กราฟทั้งหมด r_{n+1} คือ ค่า Recall จุดสุดท้าย r_n คือ ค่า Recall จุดเริ่มต้น และ Pinterp(r_{n+1}) คือ ค่า Precision สูงสุด



ภาพประกอบที่ 13 พื้นที่ใต้กราฟทั้งหมด

จากนั้นหาค่า AP_1, AP_2, AP_3 มาหาค่าเฉลี่ย จึงจะได้ค่าของ mean Average Precision (mAP) ของ Class ทั้งหมด จากการคำนวณเราจะได้ค่า mean Average Preision (mAP) = 0.176

4.วิธีการทดลอง

ในการเตรียมชุดข้อมูลได้เตรียมวิดีโอทั้งหมด 56 คลิป โดยเฉลี่ยคลิปวิดีโอละ 5-50 นาที และทำการตัดรูปภาพออกมาทุกๆ 36 วินาทีหรือ 720 เฟรม จะได้รูปภาพออกมาทั้งหมด 9,964 รูป และนำมาทำการแบ่งข้อมูลออกเพื่อใช้ตรวจสอบความถูกต้อง (Validation) ระหว่างเรียนรู้ (Train) และ ใช้ทดสอบ (Test) เพื่อวัดประสิทธิภาพแบ่งออกได้ดังนี้

ตารางที่ 4 การแบ่งข้อมูลภาพเพื่อใช้ในการตรวจจับใบหน้า

Happy	Neutral	Sad
จำนวนข้อมูลสำหรับการเรียนรู้ (80%)		จำนวนข้อมูลสำหรับวัดประสิทธิภาพ (20%)
7,971 รูป		1,993 รูป
ข้อมูลทั้งหมด 9,964 รูป		

จากชุดข้อมูลสำหรับการเรียนรู้ (Train) ที่ผ่านการทำ Data Augmentation

เพื่อเพิ่มข้อมูลจากรูปภาพใบหน้าคนจะได้รูปภาพเพิ่มขึ้นมาเป็น 15,942 รูปภาพแบ่งจำนวนอารมณ์ได้ตามตาราง

ตารางที่ 5 ชุดข้อมูลสำหรับการเรียนรู้การรู้จำใบหน้า

รหัส	ชนิด	จำนวนใบหน้า
1	Happy	11,108
2	Neutral	12,840
3	Sad	10,042
รวม		33,990

ตารางที่ 6 ชุดข้อมูลสำหรับทดสอบการรู้จำใบหน้า

รหัส	ชนิด	จำนวนใบหน้า
1	Happy	1,411
2	Neutral	1,489
3	Sad	1,239
รวม		4,163

4.1 การตั้งค่าการเรียนรู้

ตารางที่ 7 การตั้งค่าการฝึกการเรียนรู้

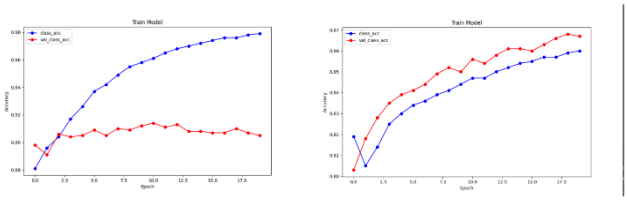
Function	Description
Epoch กำหนดการเรียนรู้	20
Train RPN ฝึกการเรียนรู้ โมเดล RPN	True
Train final classifier ฝึกการเรียนรู้ โมเดล VGG-16 ส่วนที่ทำการคัดเลือกคำตอบ	True
Train base NN ฝึกการเรียนรู้ โมเดล VGG-16 ส่วนที่หาคุณลักษณะเด่นของรูปภาพ (Feature Extraction)	True

ตารางที่ 7 การตั้งค่าการฝึกการเรียนรู้(ต่อ)

Function	Description
Anchor box scales กำหนดขนาดของ Anchor box ทั้งสามขนาด	128, 256, 512
Anchor box ratios กำหนดสัดส่วนของ Anchor box	[1 : 1], [0.7 : 1.4], [1.4 : 0.7]
Image size กำหนดขนาดภาพภาพที่ทำการฝึกทั้งหมด	300
Optimizer	Adam, Learning rate 0.0001
Augment สุ่มสร้างรูปภาพใหม่	True
Model training APIs ฟังก์ชันการฝึกการเรียนรู้ (Training)	train_on_batch

4.2 ผลการทดลอง

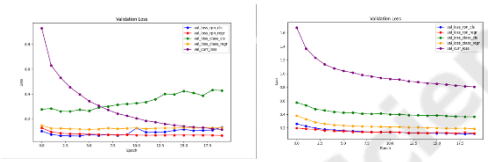
การฝึกโมเดล 20 รอบ ด้วยข้อมูลสำหรับเรียนรู้ 80 เปอร์เซ็นต์จากข้อมูลทั้งหมด และข้อมูลตรวจสอบความถูกต้องอีก 20 เปอร์เซ็นต์จากข้อมูลทั้งหมด ซึ่งการทดสอบจากข้อมูลตรวจสอบความถูกต้องได้ผลลัพธ์ที่ 0.90 ถึง 0.91 เปอร์เซ็นต์ แสดงผลการฝึกทั้ง 20 รอบด้วยรูปต่อไปนี้



ภาพประกอบที่ 14 กราฟเปรียบเทียบ

แสดงให้เห็นถึงค่า Training และ Validation ที่เกิดขึ้นจากการเรียนรู้ทั้งหมด 20

Epochs ซึ่งสามารถอธิบายได้ว่าในการเรียนรู้ในแต่ละรอบทำให้โมเดลทั้งสองตัวมีความเสถียนที่ต่างกัน โดยในส่วนของ Optimizer รูปแบบ Adam เส้นของ Training มีแนวโน้มว่าจะเพิ่มขึ้นเรื่อยๆ แต่เส้นของ Validation จะตกลงในรอบที่ 1 และเพิ่มขึ้นจนอยู่ที่ 0.90 และอิ่มตัวอยู่ที่ 0.91 โดยประมาณและมีแนวโน้มว่าจะไม่เพิ่มขึ้นมากกว่านี้ และส่วนของ Optimizer รูปแบบ SGD นั้นเส้นของ Training จะตกลงในรอบที่ 1 และเพิ่มขึ้นเรื่อยๆ แต่ค่าที่เพิ่มนั้นค่า Accuuracy ไม่สูงเท่า Adam และในส่วนของเส้น Validation จะเพิ่มขึ้นเรื่อยๆ แล้วตกในรอบที่ 9และรอบที่11



(ก) Optimizer Adam Loss (ข) Optimizer SGD Loss

ภาพประกอบที่ 15 ผลลัพธ์ค่า Loss

แสดงให้เห็นถึงค่า Loss ที่เกิดจากการเรียนรู้ทั้งหมด 20 epochs ซึ่งอธิบายได้ว่าการเรียนรู้ในแต่ละรอบนั้นทำให้โมเดลมีความเสถียรมากขึ้นโดยในส่วนของ Optimizer รูปแบบ Adam เส้น val_loss_class_cls มีแนวโน้มว่าจะเพิ่มขึ้นเรื่อยๆ เส้น val_loss_rpn_cls จะพุ่งขึ้นเล็กน้อยในช่วง 10 epochs และ Optimizer รูปแบบ SGD มีความเสถียรมากแต่ค่า Loss ของ val_curr_loss นั้นสูกมาก

ผู้จัดทำได้นำโมเดล Optimizer รูปแบบ Adam มาใช้เพราะว่าค่า Accuracy สูงกว่าส่วนค่า Loss ของ Adam นั้นมีค่า Loss ที่น้อยกว่า และ Optimizer รูปแบบ Adam มีเวลาในการ Train น้อยกว่า โดยที่ Adam ใช้เวลาทั้งหมด 20 ชั่วโมง กับอีก 30 นาที ในส่วนของ SGD ใช้เวลาทั้งสิ้น 22 ชั่วโมง กับอีก 30 นาที

4.3 ประเมินและวิเคราะห์ผล

การประเมินผลวัดจากข้อมูลสำหรับวัดประสิทธิภาพชุดทดสอบ 20 เปอร์เซ็นต์จากข้อมูลทั้งหมด และมีบางอารมณ์ ที่ไม่เท่ากันเนื่องจากทำการสุ่มข้อมูลในการทำจึงทำให้ข้อมูลบางชุดไม่เท่ากัน จากนั้นนำข้อมูลสำหรับวัดประสิทธิภาพแต่ละชนิด หาค่าความแม่นยำ (Precision), ค่าความระลึก (Recall) ด้วย IoU ที่ได้ 0.5 ขึ้นไปและถือเป็นค่าวัดที่อยู่ในระดับกลางที่ใช้กันทั่วไปในงานจรวจจับวัตถุ เมื่อได้ผลลัพธ์แล้วนำค่าทั้งสองหาค่าความแม่นยำเฉลี่ย (Average Precision)

ผลลัพธ์การตรวจจับใบหน้าทั้งหมดจะแสดงรายละเอียดการตรวจจับทั้งหมดและจำนวนผลเฉลย (Ground Truth) ตามตาราง

ตารางที่ 8 รายละเอียดการตรวจจับใบหน้า

ผลการตรวจจับใบหน้าทั้งหมด	
จำนวนใบหน้าในผลเฉลย	4,163
จำนวนใบหน้าที่ตรวจจับได้	3,255
จำนวนใบหน้าที่ทำนายถูกต้องที่ IoU > 0.5	1,948
จำนวนใบหน้าที่ทำนายผิด	1,307

จากตารางที่ 5 จะทำการบอกจำนวนผลการตรวจจับใบหน้าทั้งหมด โดยที่พบจำนวนใบหน้าในผลเฉลย 4,163 ใบหน้าจากข้อมูลชุดทดสอบ 20 เปอร์เซ็นต์ โดยมีใบหน้าที่ตรวจจับ

ได้ 3,255 ใบหน้า และมีจำนวนใบหน้าที่ทำนายถูกต้องที่ IoU > 0.5 ทั้งหมด 1,948 ใบหน้า จำนวนที่ใบหน้าทำนายผิด 1,307 ใบหน้า โดยในแต่ละใบหน้าจะมี 3 อารมณ์

ตารางที่ 9 การตรวจจับของ Happy

ผลการตรวจจับของ Happy	
จำนวนใบหน้าในผลเฉลย	1411
จำนวนใบหน้าที่ตรวจจับได้	933
จำนวนใบหน้าที่ทำนายถูกต้อง (TP)	676
จำนวนใบหน้าที่ทำนายผิดพลาด (FP)	317
ความแม่นยำเฉลี่ย (AP)	0.77

ตารางที่ 10 การตรวจจับของ Neutral

ผลการตรวจจับของ Neutral	
จำนวนใบหน้าในผลเฉลย	1513
จำนวนใบหน้าที่ตรวจจับได้	833
จำนวนใบหน้าที่ทำนายถูกต้อง (TP)	311
จำนวนใบหน้าที่ทำนายผิดพลาด (FP)	522
ความแม่นยำเฉลี่ย (AP)	0.47

ตารางที่ 11 การตรวจจับของ Sad

ผลการตรวจจับของ Sad	
จำนวนใบหน้าในผลเฉลย	1239
จำนวนใบหน้าที่ตรวจจับได้	1429
จำนวนใบหน้าที่ทำนายถูกต้อง (TP)	635
จำนวนใบหน้าที่ทำนายผิดพลาด (FP)	794
ความแม่นยำเฉลี่ย (AP)	0.54

การประเมินผลจากตารางที่ 9, 10 และ 11 ได้ผลลัพธ์ตามลำดับดังนี้ 0.77, 0.47, 0.54 เมื่อนำค่าทั้งสามชนิดมารวมกันและหาค่าเฉลี่ยทั้งหมดจะได้ค่า mAP (mean Average Precision) ดังนั้น mAP เท่ากับ 0.59 หรือ 59 เปอร์เซ็นต์ ด้วย IoU ที่ 0.5

เมื่อสรุปผลของแต่ละ Class Happy มีค่าความแม่นยำเฉลี่ยที่สูงกว่าหมวดอื่นๆ และมีความแม่นยำเฉลี่ย (AP) สูงที่สุด แปลว่าอารมณ์ Happy มีการตรวจจับวัตถุที่มีอยู่ในผลเฉลย (Ground Truth) ได้มากกว่า Class อื่นๆ อาจเป็นเพราะลักษณะหน้าสามารถมองออกได้ง่ายและมีลักษณะคล้ายๆ กัน เมื่อเทียบกับ Neutral Sad จะมีการตรวจจับวัตถุที่มีอยู่ในผลเฉลย (Ground Truth) ได้น้อยกว่า

ค่าเฉลี่ย mAP (mean Average Precision) เท่ากับ 0.59 หรือ 59 เปอร์เซ็นต์ ที่ได้ออกมา ไม่สามารถบอกได้ว่าโมเดลจะทำนายผลแต่ละ Class มากน้อยเพียงใด แต่เป็นการบอกค่าโดยรวมที่โมเดลสามารถทำนายผลออกมาได้

ตารางที่ 12 Model จำแนกอารมณ์

รายละเอียด	Happy	Neutral	Sad
AP	77%	47%	54%
ผลลัพธ์	ปกติ	ต่ำ	ปกติ
mAP	59%		

5.สรุปและอภิปรายผลการทดลอง

โครงงานปริญญานิพนธ์ฉบับนี้ได้นำเสนอโปรแกรมวิเคราะห์อารมณ์บนใบหน้าจากวิดีโอ จำแนกใบหน้าที่จะวิเคราะห์อารมณ์ได้แก่ มีความสุข (Happy), ปกติ (Neutral),

เศร้า (Sad) โดยใช้การเรียนรู้เชิงลึกด้วยวิธี Faster R-CNN (FRCNN) ซึ่งสามารถใช้งานโปรแกรมวิเคราะห์อารมณ์บนใบหน้าจากวิดีโอได้ผ่าน Desktop Application

หลังจากการพัฒนาโมเดลและทดสอบโมเดลโปรแกรมวิเคราะห์อารมณ์บนใบหน้าจากวิดีโอ ได้มีการทดสอบและวัดประสิทธิภาพโปรแกรมวิเคราะห์อารมณ์บนใบหน้าจากวิดีโอโดยใช้ชุดข้อมูลภาพใบหน้า ทั้งสิ้น 1993 ภาพ เพื่อนำมาวิเคราะห์อารมณ์บนใบหน้า พบว่าค่าความมั่นใจของแต่ละอารมณ์ได้แก่ มีความสุข (Happy) สูงสุด 0.99 ต่ำสุด 0.82, ปกติ (Neutral) สูงสุด 0.99 ต่ำสุด 0.80, เศร้า (Sad) สูงสุด 0.99 ต่ำสุด 0.82

โดยพบปัญหากับภาพที่มีตำแหน่งที่อยู่ใกล้กันมาก ถึงแม้จะไม่ใช่วัตถุเดียวกันทำให้โมเดลตรวจจับได้ยากมากขึ้นอาจเป็นเพราะกระบวนการตรวจจับวัตถุพยายามที่จะหาคุณลักษณะเด่นของภาพเมื่อวัตถุอยู่ใกล้กันมากและมีลักษณะที่ใกล้เคียงกันก็จะถูกมองว่าเป็นวัตถุเดียวกันและพลาดวัตถุนั้นไป และปัญหาอีกอย่างคือเมื่อโมเดลตรวจจับมากกว่าผลเฉลยจะทำให้ไปคำนวณค่าของ IoU ออกมาไม่ได้ IoU เป็นค่าที่ใช้ในการวัดความแม่นยำของการตรวจจับวัตถุ โดย IoU จะเป็นอัตราส่วนของพื้นที่ที่มีการซ้อนทับระหว่าง Bounding Boxes กับพื้นที่ทั้งหมดของ Bounding Boxes ทั้งสอง ค่า IoU จะอยู่ในช่วง 0-1 โดยค่า 1 หมายถึง Bounding Boxes สองชิ้นมีส่วนทับซ้อนกันทั้งหมด ส่วนค่าต่ำกว่านั้นจะหมายถึงว่าการตรวจจับไม่แม่นยำเท่าที่ควรเป็น โดยทั่วไปแล้วค่า IoU ที่มากกว่าหรือเท่ากับ 0.5 ถือว่าดีสำหรับการตรวจจับวัตถุโดยทั่วไป

จากการใช้งานจริงด้วยวิดีโอที่โหลดมาลองใช้ได้ผลลัพธ์ที่ไม่ดีมากนัก อาจเป็นเพราะโมเดลมองภาพนั้นเป็น ใบหน้าเศร้า แต่อ้าปากเห็นฟัน โมเดลอาจทำนายผลออกมา มีความสุข (Happy) ก็เป็นได้ ที่เป็นแบบนี้อาจเป็นเพราะ Dataset ภาพแบบนั้นน้อยเกินไปแต่ละอารมณ์ การวิเคราะห์อาการอารมณ์ของมนุษย์เป็นเรื่องที่ซับซ้อนและมีความยุ่งยากอย่างมาก ไม่ว่าจะเป็นการวิเคราะห์ในเชิงพฤติกรรมหรือพฤติกรรมทางกายภาพ หรือการวิเคราะห์ความรู้สึกภายในเพื่อหาอาการอารมณ์ เช่น ความสุข ความเศร้า ความโกรธ และอื่น ๆ นั้นมีความยุ่งยากเพราะผลลัพธ์ของการวิเคราะห์อาการอารมณ์ไม่ใช่สิ่งที่เป็นที่แน่นอนเสมอไป อาการอารมณ์ของมนุษย์มีความซับซ้อนและมีลักษณะที่แตกต่างกันไปในแต่ละบุคคล นอกจากนี้ การวิเคราะห์อาการอารมณ์ยังต้องพิจารณาปัจจัยหลายอย่าง เช่น สถานการณ์ที่เกิดขึ้น ประสบการณ์ที่ผ่านมา และภูมิสภาพทางจิตใจในขณะนั้น ดังนั้น การวิเคราะห์อาการอารมณ์ของมนุษย์ไม่ใช่เรื่องที่ง่ายและมีความซับซ้อนอย่างมาก

5.1 ปัญหาและอุปสรรคในการดำเนินงาน

5.1.1 อัลกอริทึมการตรวจจับ Faster R-CNN และ VGG16 มีความซับซ้อนสูงมาก จำเป็นต้องใช้ทรัพยากรในการประมวลผลสูงและใช้ระยะเวลานาน

5.1.2 ชุดข้อมูลที่ใช้ในการเรียนรู้มีไม่เพียงพอและไม่มีความหลากหลาย อย่างเช่น เราเห็นใบหน้าเป็นอารมณ์มีความสุข แต่โมเดลอาจจะมองเป็นอารมณ์เศร้าก็ได้ แล้วความรู้สึกของแต่ละคนก็ไม่เหมือนกันว่าจะมองเป็นอารมณ์แบบไหน

5.1.3 อัลกอริทึมการตรวจจับวัตถุจำเป็นต้องมีการทำผลเฉลยด้วยตัวเองและในหนึ่งภาพอาจมีได้หลายวัตถุจึงใช้ระยะเวลานานในการสร้างผลเฉลย แต่บางภาพที่โมเดลทำนายผลออกมาก็ตรวจจับได้หลายวัตถุ จึงทำให้ภาพที่เป็นผลเฉลยมีวัตถุไม่เพียงพอเลยหาค่า IoU ออกมาไม่ได้

5.2 ข้อเสนอแนะ

ควรมีชุดข้อมูลจากหลากหลายแหล่งที่มีลักษณะของข้อมูลที่แตกต่างกัน เพื่อให้เกิดชุดข้อมูลที่มีความหลากหลาย

เอกสารอ้างอิง

1.J. Jayalekshmi and T. Mathew, " Facial expression recognition and emotion classification system for sentiment analysis”,2017.[Online].Available: https://ieeexplore.ieee.org/abstract/document/8076732?casa_token=cW7UK2ACD8YAAAAA:kp3F5lBawGoEGmJeJGYAE5BBoeyzsYx0uc8YSZa6H6V13fMLF957K6OLLrtCjD8OEEL_l90Vjhj.

2.จุฑามาศ มาบรรดิษ, "การรู้จำอารมณ์บนใบหน้า โดยใช้วิธีการวิเคราะห์องค์ประกอบหลัก และการวิเคราะห์จำแนกประเภทเชิงเส้น”, 2014.[Online].Available: https://www.eng.kps.ku.ac.th/dblibv2/fileupload/project_IdDoc46_IdPro457.pdf.

3.ขวัญชัย กรอนันต์ศิลป์, "ระบบจดจำใบหน้า (Face Recognition) เป็นเทคโนโลยีที่มาช่วยในการยืนยันอัตลักษณ์ของบุคคล ซึ่งหากภาครัฐนำระบบจดจำใบหน้ามาประยุกต์ใช้ในเรื่องความปลอดภัยส่วนบุคคล”, Feb 2021.[Online].Available: https://archive.cm.mahidol.ac.th/handle/123456789/3855.

4.T.F. Cootes และ C.J. Taylor, " Locating faces using statistical feature detectors”, Oct 1996.[Online].Available: https://ieeexplore.ieee.org/document/557265

5.ธนพล พุ่มจำเจียก, "การรู้จำอารมณ์บใบหน้าจากวิดีโอ โดยใช้ตัวกรองกาบอร์ วิธีกาวิเคราะห์องค์ประกอบหลักและการวิเคราะห์จำแนกประเภทเชิงเส้น”,2016.[Online].Available: https://www.it.kmitl.ac.th/~sirion/senior_project/facial_expression/55070053.pdf

6.พุฒิพงศ์ จันทร์แจ่ม, "การปรับปรุงประสิทธิภาพของกระบวนการบริการลูกค้าโดยใช้การวิเคราะห์จากกล้องวรจรปิด”, 2020.[Online].Available: https://repository.nida.ac.th/bitstream/handle/662723737/5540/6110412029.pdf?sequence=1

7.Natthasath Saksupanara, "Using the Deep Learing for Garbage Detection with the Applied of Smart Bin” , Sep 20, 2019.[Online].Available: https://www.slideshare.net/FSDotNet/deep-learning-smart-bin

8.Sanparith Marukatat, "โลกหมุนไปงานวิจัยก็หมุนตาม”, May 8, 2018.[Online].Available: https://www.slideshare.net/FSDotNet/deep-learning-smart-bin