การจำแนกอารมณ์จากใบหน้าแบบ Real-time บนอุปกรณ์ฝังตัว ด้วยแบบจำลองการเรียนรู้เชิงลึก

ฐิติพงษ์ รักษาริกรณ์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่
วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์
มหาวิทยาลัยธุรกิจบัณฑิตย์
พ.ศ. 2563

Real-time Face Expression Classification on Embedded Devices using Compact Deep Learning Model

Thitiphong Raksarikorn

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Engineering Department of Big Data Engineering, College of innovative Technology and Engineering, Dhurakij Pundit University

2020

ใบรับรองงานวิทยานิพนธ์

วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์

ปริญญา วิศวกรรมศาสตรมหาบัณฑิต

หัวข้อวิทยานิพนธ์ การจำแนกอารมณ์จากใบหน้าแบบ Real-time บนอุปกรณ์ฝังตัวด้วยแบบจำลองการเรียนรู้เชิงลึก

เสนอโดย ฐิติพงษ์ รักษาริกรณ์

สาขาวิชา วิศวกรรมข้อมูลขนาดใหญ่

อาจารย์ที่ปรึกษาวิทยานิพนธ์ ดร.ธนภัทร ฆังคะจิตร

ได้พิจารณาเห็นชอบโดยคณะกรรมการสอบวิทยานิพนธ์แล้ว

.....ประธานกรรมการ

(รองศาสตราจารย์ ดร.กฤษณะ ไวยมัย)

.....กรรมการและอาจารย์ที่ปรึกษาหลัก

(ดร.ธนภัทร ฆังคะจิตร)

.....กรรมการ

(ดร.สรรพฤทธิ์ มฤคทัต)

.....กรรมการ

(ดร.เอกสิทธิ์ พัชรวงศ์ศักดา)

วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์รับรองแล้ว

.....

(ผู้ช่วยศาสตราจารย์ ดร.ณรงค์เดช กีรติพรานนท์)

คณบดีวิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์

วันที่ 17 เดือน กรกฎาคม พ.ศ. 2563

หัวข้อวิทยานิพนธ์ การจำแนกอารมณ์จากใบหน้าแบบ Real-time บนอุปกรณ์ฝังตัวด้วยแบบจำลองการเรียนรู้เชิงลึก

ชื่อผู้เขียน ฐิติพงษ์ รักษาริกรณ์

อาจารย์ที่ปรึกษา ดร.ธนภัทร ฆังคะจิตร

สาขาวิชา วิศวกรรมข้อมูลขนาดใหญ่

ปีการศึกษา 2562

บทคัดย่อ

การแสดงอารมณ์ผ่านทางสีหน้าถือเป็นรูปแบบสำคัญของการสื่อสารในชีวิตประจำวันของมนุษย์ การทำความเข้าใจอารมณ์เหล่านี้ทำให้สามารถเข้าใจผู้คนรอบข้างได้มากขึ้น ดังนั้นการจำแนกการแสดงออกทางสีหน้าโดยอัตโนมัติจึงเป็นปัญหาสำคัญอย่างหนึ่งในงานวิจัยด้านคอมพิวเตอร์วิทัศน์ งานวิจัยสมัยใหม่ทางด้านนี้ได้ประยุกต์ใช้เทคนิคการเรียนรู้เชิงลึกเพื่อหลีกเลี่ยงกระบวนการแยกคุณลักษณะที่ซับซ้อนและให้ความแม่นยำสูง อย่างไรก็ตามโดยส่วนใหญ่ของโครงข่ายประสาทเทียมแบบคอนโวลูชั่น (CNNs) ที่มีประสิทธิภาพสูง (เช่น XCEPTION) มักจะมีขนาดใหญ่และซับซ้อนทำให้ไม่สามารถนำไปใช้บนอุปกรณ์ฝังตัว ถึงแม้ว่าจะมีโครงข่ายประสาทเทียมบางประเภทที่มีขนาดเล็ก (เช่น MobileNet) แต่ให้ความแม่นยำในการทำนายที่ค่อนข้างต่ำ ดังนั้นในงานวิจัยนี้จึงได้เสนอสถาปัตยกรรมโครงข่ายประสาทเทียมแบบคอนโวลูชั่นที่มีขนาดเล็กพอที่จะทำงานบนอุปกรณ์ฝังตัวและให้ความแม่นยำสูงในการจำแนกการแสดงออกทางสีหน้า โดยโครงข่ายประสาทเทียมที่นำเสนอมีพารามิเตอร์เพียง 2.2 ล้านเท่านั้นซึ่งน้อยกว่า XCEPTION ประมาณ 10 เท่า และมีความสามารถในการทำงานบนอุปกรณ์ฝังตัวเช่นเดียวกับ MobileNet ได้

ผลการทดลองกับชุดข้อมูลมาตรฐาน (FER-2013) แสดงให้เห็นว่าโครงข่ายประสาทเทียมที่นำเสนอให้ความแม่นยำเทียบเท่า (0.7169) กับ XCEPTION และมีความแม่นยำมากกว่าระดับการจำแนกโดยใช้มนุษย์ (0.65±5) รวมถึงจากการทดสอบการจำแนกใบหน้าบนอุปกรณ์ฝังตัว (Raspberry Pi) สำหรับโมเดลจำแนกอารมณ์นักเรียนโดยการจับภาพจากกล้องเว็บแคมในห้องเรียน 1 ภาพต่อ 1 และ 2 วินาที พบว่ามีความเร็วในการประมวลผลใกล้เคียงกับ MobileNet โดยใช้เวลาน้อยกว่า 1 วินาที ซึ่งเพียงพอต่อการทำงานแบบ Real-time ของการจำแนกการแสดงอารมณ์บนอุปกรณ์ฝังตัว

คำสำคัญ: อารมณ์, การจำแนกการแสดงออกทางสีหน้า, โครงข่ายประสาทเทียมแบบคอนโวลูชั่นขนาดเล็ก, อุปกรณ์ฝังตัว, การจำแนกแบบ Real-time

Thesis Title	Real-time face expression classification on embedded devices Using compact deep learning model
Author	Thitiphong Raksarikorn
Thesis Advisor	Dr.Thanapat Kangkachit
Department	Big Data Engineering
Academic Year	2019

ABSTRACT

Facial expressions play crucial role in communication in human life. Learning to read these emotions is incredibly helpful for better understanding the people in our lives. Therefore, automated facial-expression classification is an essential research topic in computer-vision field. Recently, many researches applied deep learning techniques to avoid complex feature extraction process and to obtain satisfied classification performance. However, most of the effective convolutional neural network (CNN) architectures (e.g. XCEPTION) are complicated and contain a large number of parameters which can not be fit in embedded devices. Although, there existed many compact CNNs architectures (e.g. MobileNet), most of them provided unsastified classification performance. In this work, we propose a compact CNN architecture that is capable of working embedded devices and effectively classify facial expressions. To efficiently use of model parameters, our proposed architecture has only 2.2 million parameters which is about 10 times less than XCEPTION. Meanwhile, our proposed architecture is capable of working on embedded devices as the same as MobileNet.

The experimental results on FER-2013 dataset show that our model offers comparable accuracy (0.7169) to XCEPTION and the upper-bound level of human accuracy (0.65±5). To test runtime efficiency of our moel on embedded devices (Raspberry Pi), the experiments were conducted to classify student facial-expressions in a real classroom. An image for every 1 and 2 seconds were captured from realtime streaming video. As result, our model offers comparable runtime to MobileNet (i.e. less than 1 second) which adequate for the needs in real-time classifying facial-expressions on embedded devices.

keyword: emotions, facial expression classification, compact convolutional neural networks, embedded devices, real-time classification

กิตติกรรมประกาศ

งานวิจัยนี้สำเร็จได้ด้วยการสนับสนุนและความช่วยเหลือจากหลายส่วน ขอบพระคุณที่ปรึกษาวิทยานิพนธ์ ดร. ธนภัทร ฆังคะจิตร ที่ให้คำแนะนำ เสนอแนะ และผลักดันให้งานวิจัยนี้สำเร็จลุล่วงไปได้ ขอบคุณคณะกรรมการสอบวิทยานิพนธ์ที่สละเวลาให้คำแนะนำและแนวทางให้งานวิจัยนี้สำเร็จ ขอบคุณอาจารย์และเจ้าหน้าที่หลักสูตรวิศวกรรมศาสตร์มหาบัณฑิต สาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่ทุกท่านที่ให้ความรู้คำปรึกษาและประสานงานให้งานวิจัยเป็นไปอย่างราบรื่น ขอบพระคุณ ผู้ช่วยศาสตร์จารย์ ดร.วรพล พงษ์เพชร ที่ให้คำแนะนำในเรื่อง การจำแนกอารมณ์ทางใบหน้า และขอบคุณ ดร.เอกสิทธิ์ พัชรวงศ์ศักดา ที่สนับสนุนเป็นอย่างดีมาตลอดในทุกเรื่อง ขอบคุณเพื่อนทุกท่านที่ให้คำปรึกษา ขอบคุณครอบครัวที่เป็นกำลังใจ ช่วยเหลือและให้การสนับสนุนในทุกๆ เรื่อง ผู้วิจัยหวังเป็นอย่างยิ่งว่าวิทยานิพนธ์ฉบับนี้จะเป็นพื้นฐานในการต่อยอดองค์ความรู้ของผู้ที่สนใจศึกษาในงานด้านนี้ต่อไป

ฐิติพงษ์ รักษาริกรณ์

สารบัญ

สารบัญ (ต่อ)

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญภาพ

ภาพที่ **หน้า**

สารบัญภาพ (ต่อ)

สารบัญภาพ (ต่อ)

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ในปัจจุบันการสื่อสารเป็นเรื่องที่สำคัญของมนุษย์ ซึ่งในส่วนใหญ่มนุษย์มักแสดงอารมณ์ออกทางสีหน้าเป็นส่วนสำคัญ ในการแสดงอารมณ์ออกทางสีหน้าเป็นรูปแบบของการสื่อสารในรูปแบบอวัจนภาษา ซึ่งทำให้สามารถตีความได้เร็วกว่าการสื่อสารในรูปแบบวัจนภาษา หรือ วาจา และสามารถตรวจสอบอารมณ์ของมนุษย์ได้จากการแสดงออกทางสีหน้าในระหว่างการปฏิสัมพันธ์ระหว่างมนุษย์ทำให้ทราบถึงอารมณ์ของผู้ร่วมสนทนา คำนิยามการแสดงออกทางสีหน้าที่มีชื่อเสียงและใช้กันอย่างแพร่หลายถูกกำหนดไว้ใน)M. Harinthapatip,2017) ที่เรียกว่า FACS (Facial Action Coding System) ซึ่งใน FACS ได้กำหนดกลุ่มของการแสดงออกทางสีหน้า 7 กลุ่ม ได้แก่ ความสุข (Happiness), เศร้า (Sad), ความโกรธ (Anger), ความกลัว (Fear), ความแปลกใจ (Surprise), รังเกียจ (Disgust), และดูถูก (Contempt) ซึ่งวิธีการอัตโนมัติในการจดจำการแสดงออกทางสีหน้าเหล่านี้เป็นมิติใหม่สำหรับปฏิสัมพันธ์ระหว่างมนุษย์กับคอมพิวเตอร์ ตัวอย่างของแอพพลิเคชันดังกล่าว คือ ความพึงพอใจของลูกค้าในระหว่างการรับบริการ ความตั้งใจของนักเรียนในห้องเรียน ปฏิกิริยาที่เหมาะสมโดยใช้หุ่นยนต์ (P. Ekman, 1934)

การจำแนกประเภทของการแสดงอารมณ์ออกทางสีหน้ากลายเป็นปัญหาสำคัญในสาขาการวิจัยเกี่ยวกับด้านคอมพิวเตอร์วิทัศน์)Computer Vision: CV) ต่อมาได้มีการพัฒนากระบวนการการวิจัยที่เกี่ยวข้องกับกับการจัดหมวดหมู่ภาพโดยใช้โครงข่ายประสาทเทียมแบบคอนโวโลชั่น)Convolutional Neural Networks; CNNs((S. Alizadeh and A. Fazel, 2017) ซึ่งถูกใช้กันอย่างแพร่หลายเนื่องจากมีความแม่นยำในการทำนายสูงและมีกระบวนการสกัดคุณลักษณะสำคัญจากภาพนำเข้า โดยในช่วงต้นของโครงข่าย CNNs จะทำการสกัดคุณลักษณะที่หยาบ ในขณะที่คุณลักษณะที่ซับซ้อนมากขึ้นจะถูกสกัดในชั้นลึกของโครงข่าย อย่างไรก็ตามโครงข่าย ที่มี CNNs ประสิทธิภาพในการทำนายสูงเป็นโครงข่ายที่มีความซับซ้อนมากและมีขนาดใหญ่ทำให้ไม่สามารถใช้งานในอุปกรณ์ฝังตัว (Embedded Devices)

ด้วยเหตุนี้ Google จึงได้มีการพัฒนา MobileNet [Natalia Efremova, Mikhail Patkin and Denis Sokolov ,2019] ซึ่งเป็นโครงข่ายที่ถูกออกแบบมาสำหรับมือถือ โดยเป็นโครงข่ายที่มีขนาด

เล็กและไม่ซ้อนซ้อนมาก ทำให้สามารถใช้งานโครงข่ายได้โดยไม่ใช้ทรัพยากรมากนัก แต่มีข้อเสียตรงประสิทธิภาพในการทำนายที่มีความแม่นยำน้อยลงเมื่อเทียบกับโครงข่าย CNNS อื่น เช่น Inception[Bharath Raj, 2018], Resnet[Priya Dwivedi, 2019], XCEPTION [Octavio Arriaga and Paul G. Ploger, 2017] เป็นต้น

การตรวจจับอารมณ์แบบreal-time จากกล้องดิจิตอลที่ทำการส่งวิดีโอผ่านระบบเครือข่ายเน็ตเวิร์ก)Video Streaming) สามารถรายงานผลได้ทันทีและตลอดเวลา แต่มีข้อเสียคือมีปริมาณข้อมูลจำนวนมากไหลผ่านระบบเครือข่าย ทำให้อาจเกิดปัญหาความคับคั่งของปริมาณข้อมูลในระบบเครือข่ายได้ ดังนั้นจึงจำเป็นต้องใช้อุปกรณ์ฝังตัว (Rasberry Pi) ในการช่วยประมวลผลข้อมูลวิดีโอจากกล้องดิจิตอล แล้วจึงทำการส่งผลลัพธ์จากการประมวลผลไปในระบบเครือข่าย ซึ่งทำให้สามารถลดปัญหาความคับคั่งของปริมาณข้อมูลในระบบเครือข่ายได้ อย่างไรก็ตามเนื่องจากข้อจำกัดทางด้านทรัพยากรฮาร์แวร์ของ Raspberry Pi ทั้งในด้านหน่วยประมวลผลและหน่วยความจำ จึงจำเป็นต้องพัฒนาและปรับปรุงโมเดลให้มีขนาดเล็กลง ในขณะที่ยังคงความถูกต้องแม่นยำในระดับหนึ่ง)Octavio Arriaga and Paul G. Ploger, 2017) Raspberry Pi เป็นที่นิยมในการเขียนคำสั่งโปรแกรมเพื่อสั่งการอุปกรณ์ต่างๆ เนื่องจากมีระบบปฏิบัติการที่รองรับการทำงานการเขียนโปรแกรม

ดังนั้นในงานวิจัยนี้ได้เสนอสถาปัตยกรรมโครงข่ายประสาทเทียมแบบคอนโวลูชั่นที่มีขนาดเล็กพอที่จะทำงานบนอุปกรณ์ฝังตัวและให้ความแม่นยำสูงในการจำแนกการแสดงออกทางสีหน้า โดยโครงข่ายที่นำเสนอมีจำนวนพารามิเตอร์เพียง 2.ล้านเท่านั้นซึ่งน้อยกว 2่าโครงข่ายต้นแบบ XCEPTION ประมาณ เท่า และให้ความแม่นยำในการทำนายใกล้เคียงกันบนชุดข้อมูล 10 มาตรฐานFER-2013 Challenges in Representation Learning ICML. (2013) นอกจากนี้โครงข่ายที่นำเสนอสามารถประมวลผลบนอุปกรณ์ฝังตัวขนาดเล็กได้ เช่น)Raspberry Pi) โดยใช้เวลาในการประมวลผลใกล้เคียงกับโครงข่ายขนาดเล็กอย่าง MobileNet

1.2 วัตถุประสงค์ของงานวิจัย

เพื่อพัฒนาระบบจำแนกอารมณ์จากใบหน้าแบบ Real-time โดยใช้ โครงสร้างเครือข่ายประสาทเทียม (Convolutional Neural Networks: CNNs)

1.2.1 มีประสิทธิภาพเทียบเคียงกับโมเดล state-of-art: Xception Model

1.2.2 โมเดลที่ได้มีขนาดเล็กพอที่จะสามารถใช้งานบน Raspberry Pi ได้

1.2.3 การรับข้อมูล (frame-rate) เข้ามาประมวลผลเพียงพอต่อการทำงานของระบบ

1.3 ขอบเขตงานวิจัย

1.3.1 ขอบเขตข้อมูล

จำแนกตามการจัดกลุ่มของใบหน้าที่มีการแสดงอารมณ์ออกทางสีหน้า ได้ทั้งหมด 7 กลุ่ม ดังนี้ 1. โกรธ (Angry) 2. รังเกียจ (Disgust) 3. กลัว (Fear) 4. มีความสุข (Happy) 5. เศร้า (Sad) 6. แปลกใจ (Surprise) และ 7. เป็นกลาง (Neutral) ตามลำดับ

ภาพใบหน้าแสดงอารมณ์จาก www.kaggle.com คือ dataset FER2013 (Challenges in Representation Learning ICML, 2013) ที่เป็นข้อมูลกลาง ประกอบด้วยข้อมูลภาพสีเทา มีขนาด 48x48 พิกเซล โดยมีการจัดกลุ่มของใบหน้าที่มีการแสดงอารมณ์ออกทางสีหน้า ได้ทั้งหมด 7 กลุ่ม ดังนี้ 1. โกรธ 2. รังเกียจ 3. กลัว 4. มีความสุข 5. เศร้า 6. แปลกใจ และ 7. เป็นกลาง ตามลำดับ มีจำนวน 35,887 ใบหน้า

1.4 สมมติฐานของงานวิจัย

ได้ระบบจำแนกอารมณ์จากใบหน้าแบบ Real-time โดยใช้ โครงสร้างเครือข่ายประสาทเทียม)Convolutional Neural Networks: CNNs) ที่มีประสิทธิภาพเทียบเคียงกับโมเดล state-of-art: Xception Model ที่สามารถใช้งานบน Raspberry Pi ได้

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1.5.1. อารมณ์ หมายถึง เป็นการแสดงออกที่เกิดมาจากสภาวะทางจิตใจของมนุษย์ ที่ได้รับผลกระทบหรือสิ่งเร้ามากระตุ้นให้เกิดปฏิกิริกยาการตอบสนองออกมา ซึ่งโดยทั่วไป "อารมณ์" จะส่งผลต่อวิธีการคิดและส่งผลต่อการดำเนินชีวิตประจำวัน รวมถึงการตัดสินใจ ที่ส่งผลต่อความเข้าใจและการใช้อารมณ์ต่างๆ ได้อย่างเหมาะสม จึงจำเป็นที่จะต้องทำความเข้าใจอารมณ์อย่างเหมาะสมในแต่ละสถานการณ์

1.5.2. การแสดงออกทางสีหน้าของมนุษย์ หมายถึง การจัดประเภทอารมณ์ที่แสดงออกทางสีหน้าของมนุษย์ เช่น อารมณ์โกรธจะเริ่มจากความรู้สึกรำคาญใจ ไปจนถึงโกรธเกรี้ยวเดือดดาล หรืออารมณ์กลัวจะมีความรุนแรง นอกจากนี้เรายังมีอารมณ์หลายๆ อารมณ์ในเวลาเดียวกันจนแยกไม่ออกว่าเป็นอารมณ์อะไรบ้าง บอกได้แต่เพียงว่า รู้สึกไม่สบายใจ เช่น ความวิตกกังวลซึ่งเป็นความรู้สึกที่คละเคล้าปนเปกันระหว่างความกลัว โกรธ เศร้า สำนึกผิด อับอาย เป็นต้น

1.5.3. การวิเคราะห์ข้อมูลจากภาพ หมายถึง การนำภาพถ่ายมาวิเคราะห์ เพื่อหาสิ่งต่างๆ ที่ซ้อนอยู่ในภาพ โดยสามารถดูค่าของสีแต่ละ pixel เพื่อนำมาใช้จำแนกความเหมือนความแตกต่างของรูปภาพนั้นๆ

1.5.4. ระบบฝังตัว หรือ สมองกลฝังตัว (embedded system) หมายถึง ระบบประมวลผล ที่ใช้ชิปหรือไมโครโพรเซสเซอร์ ซึ่งเป็นระบบคอมพิวเตอร์ขนาดจิ๋วที่ฝังไว้ในอุปกรณ์อิเล็กทรอนิกส์ เช่น Raspberry Pi ทำให้มีระบบประมวลผลที่เหมือนเครื่องคอมพิวเตอร์ทั่วไป

บทที่ 2
แนวคิด ทฤษฎี และผลงานวิจัยที่เกี่ยวข้อง

วิจัยเรื่องนี้มีวัตถุประสงค์เพื่อจำแนกอารมณ์จากใบหน้าแบบ Real-time โดยใช้ โครงสร้างเครือข่ายประสาทเทียม (Convolutional Neural Networks: CNNs) ที่มีประสิทธิภาพเทียบเคียงกับโมเดล state-of-art: Xception Model และโมเดลที่ได้มีขนาดเล็กพอที่จะสามารถใช้งานบน Raspberry Pi ได้ โดยจำเป็นต้องศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง ดังรายการต่อไปนี้

1. อารมณ์และการแสดงออกทางสีหน้าของมนุษย์ (Emotion & Facial Expression)
   - 1.1 อารมณ์ของมนุษย์
   - 1.2 การจัดประเภทอารมณ์ที่แสดงออกทางสีหน้าของมนุษย์
2. สถาปัตยกรรมเครือข่ายประสาทเทียม (Neural Network Architectures)
   - 2.1 โครงข่ายประสาทเทียมแบบ Multilayer Perceptron
   - 2.2 ความรู้พื้นฐานเกี่ยวกับ CNNs
   - 2.3 สถาปัตยกรรมนิวรอลเน็ตเวิร์กเชิงลึก CNNs Architecture
     - 2.3.1 VGG
     - 2.3.2 Inception V3, V2
     - 2.3.3 ResNet
     - 2.3.4 Xception
     - 2.3.5 E-Net
     - 2.3.6 MobileNet
3. ศึกษาเทคนิคการวิเคราะห์ใบหน้าแสดงอารมณ์
   - 3.1 SIFT-based (Scale Invariant Feature Transform)
   - 3.2 CNN-based (Convolutional Neural Networks)

4. การวัดประสิทธิภาพ (Performance Evaluation)

5. งานวิจัยที่เกี่ยวข้อง

2.1. อารมณ์และการแสดงออกทางสีหน้าของมนุษย์ (Emotion & Facial Expression)

อารมณ์ (Emotion) หมายถึง สภาวะทางจิตใจที่เกิดความปั่นป่วน ตื่นเต้น หรือเปลี่ยนแปลง เมื่อมีสิ่งเร้ามากระตุ้นซึ่งจะเกิดขึ้นอย่างฉับพลันทันที โดยเราจะไม่สามารถสังเกตเห็นได้โดยตรง แต่จะสังเกตเห็นได้ทางอ้อม โดยดูจากการเปลี่ยนแปลงทางด้านพฤติกรรมต่างๆ ที่มิได้แสดงออกมา เป็นคำพูด(Non-Verbal Behavior) เช่น การแสดงออกทางสีหน้า กิริยาท่าทาง เป็นต้น การแสดงออกทางอารมณ์ที่มีต่อประสบการณ์ในรูปแบบที่สังคมยอมรับ แต่ยืดหยุ่นพอที่จะเป็นปฏิกิริยาตอบสนองแบบฉับพลันของมนุษย์

1. อารมณ์ของมนุษย์

อารมณ์ มาจากภาษาอังกฤษ "Emotion" มีความหมายว่าการเกิดการเคลื่อนไหว หรือภาวะที่ตื่นเต้น มันเป็นการยากที่จะบอกว่า อารมณ์คืออะไร แต่มีแนวคิดหนึ่ง ที่ให้ความเข้าใจได้ง่ายกล่าวไว้ว่า อารมณ์เป็นความรู้สึกภายในที่เร้า ให้บุคคลกระทำ หรือเปลี่ยนแปลงภายในตัว ของเขาเอง ซึ่งความรู้สึกเหล่านี้จะเป็นความรู้สึกที่พึงพอใจ ไม่พึงพอใจ หรือรวมกันทั้งสองกรณี อารมณ์เป็นสิ่งที่ไม่คงที่มีการแปรเปลี่ยน อยู่ตลอดเวลา จากความหมายและธรรมชาติของอารมณ์ ทำให้นักจิตวิทยาทั้งหลายมีความเห็นว่าองค์ประกอบของอารมณ์จะแบ่งออกเป็น 3 อย่าง (Baron, 1990) ดังนี้

สภาวะการรู้คิด (cognitive states) เป็นความรู้สึกของผู้ที่กระทำหรือประสบการณ์ต่าง ๆ ของบุคคล อย่างเช่น เราเคยรู้สึกโกรธ ร่าเริง สะอิดสะเอียน เป็นต้น

ปฏิกิริยาทางสรีระ (physiological reactions) เป็นการเกิดการเปลี่ยนแปลงภายในร่างกายของเรา เช่น หัวใจเต้นเร็วขึ้นเมื่อรู้สึกตื่นเต้นหรือตกใจ

การแสดงออกของพฤติกรรม (expressive bahaviors) เป็นสัญญาณการแสดงออกของสภาวะภายใน เช่น เกิดความพอใจก็จะแสดงการยิ้ม หรือเมื่อโกรธก็อาจกล่าววาจาต่อว่าออกมา หรือแสดงการกระทืบเท้า, ตบตี เป็นต้น

อารมณ์มีอยู่มากมายหลายชนิดซึ่งเราอาจเรียกมันว่าอะไรก็ตาม แต่ว่าอารมณ์เหล่านั้น ก็มีความเด่นชัดและเป็นอิสระ นักจิตวิทยาได้จำแนก อารมณ์ โดยคำนึง สิ่งเร้าที่มาเป็นตัวกระตุ้น และรูปแบบการตอบสนองพฤติกรรมที่มีต่อสิ่งเร้านั้น และส่วนมากมีความเชื่อว่า บุคคลมีอารมณ์พื้นฐานอยู่ 3 ชนิด คือ ความโกรธ (anger) ความกลัว (fear) และความพึงพอใจ (pleasure) (Nelson ,J.L., Carlson K. And Palonsky, S.B., 1993) ส่วนอารมณ์ อื่นๆ เป็นผลที่เกิดจากอารมณ์ใดอารมณ์หนึ่ง หรือมากกว่าของอารมณ์ทั้งสามนี้ ตัวอย่างเช่นรังเกียจ เดือดดาล เครียดแค้น เป็นรูปแบบของอารมณ์โกรธ การอิจฉาและความรู้สึกผิดจะอยู่บนพื้นฐานของความกลัว ความรักและความสุขจะมีพื้นฐานมาจากความรู้สึกพึงพอใจ ความโศกเศร้าเป็นเสมือนการรวมกันของอารมณ์กลัวและอารมณ์โกรธ ทุกคนเคยมีอารมณ์โกรธ กลัว และพึงพอใจมาแล้ว แต่ทั้งอารมณ์โกรธ กลัว และพึงพอใจ เกิดมาจากสาเหตุที่แยกออกได้แตกต่างกัน ซึ่งมีความเป็นไปได้ที่เราจะจัดการหรือควบคุมมัน

พรรณทิพย์ ศิริวรรณุศย์ (2530) อารมณ์เป็นภาพการเปลี่ยนแปลงของร่างกายและจิตใจอันเนื่องมาจากปฏิสัมพันธ์ระหว่างสิ่งเร้าและอินทรีย์ และการแสดงออกโต้ตอบนั้นเป็นไปตามสถานการณ์

จึงสามารถสรุปได้ว่า อารมณ์ เป็นการแสดงออกที่เกิดมาจากสภาวะทางจิตใจของมนุษย์ ที่ได้รับผลกระทบหรือสิ่งเร้ามากระตุ้นให้เกิดปฏิกิรกยาการตอบสนองออกมา ซึ่งโดยทั่วไป "อารมณ์" จะส่งผลต่อวิธีการคิดและส่งผลต่อการดำเนินชีวิตประจำวัน รวมถึงการตัดสินใจ ที่ส่งผลต่อความเข้าใจและการใช้อารมณ์ต่างๆ ได้อย่างเหมาะสม จึงจำเป็นที่จะต้องทำความเข้าใจอารมณ์อย่างเหมาะสมในแต่ละสถานการณ์

2. การจัดประเภทอารมณ์ที่แสดงออกทางสีหน้าของมนุษย์

การแสดงความรู้สึกทางใบหน้าจัดว่าเป็นสิ่งที่สำคัญอย่างยิ่ง เพราะเป็นการแสดงออกถึงสภาวะทางอารมณ์ต่างๆ ในตัวเราและบุคคลอื่นที่สามารถสังเกตเห็นได้บ่อยๆ นักสรีรวิทยาที่ว่า ใบหน้าสามารถแสดงความรู้สึกได้แตกต่างกัน

P. Ekman, 1934 นักจิตวิทยาและผู้บุกเบิกที่โด่งดังด้านการวิจัยอารมณ์ของมนุษย์จากการแสดงออกทางสีหน้า (Facial Expression) และยังเป็นเจ้าของผลงานวิจัยที่มีชื่อเสียง Facial Action Coding System (FACS) โดยได้ระบุอารมณ์ออกเป็น 7 ประเภท ได้แก่

1. ความสุข (Happiness) เมื่อเรารู้สึกมีความสุข กล้ามเนื้อแก้มจะถูกยกขึ้น และกล้ามเนื้อรอบดวงตาหดตัว เกิดเป็นรอยย่นที่หางตา มุมปากโค้งขึ้นเป็นรอยยิ้มสมมาตร รอยย่นรอบดวงตาเป็นจุดสำคัญที่ใช้แยกรอยยิ้มที่จริงใจออกจากรอยยิ้มแสแสร้ง

2. ความเศร้า (Sadness) ในขณะที่มีอารมณ์เศร้า หัวคิ้วของเราจะย่นเข้าหากัน ทำให้เกิดริ้วย่นรูปตัวยู่คว่ำ และเกิดร่องแนวตั้งที่หว่างคิ้ว ทั้งหมดนี้เรียกว่า "กล้ามเนื้อเศร้าของดาร์วิน" มุมปากด้านนอกจะตกลง

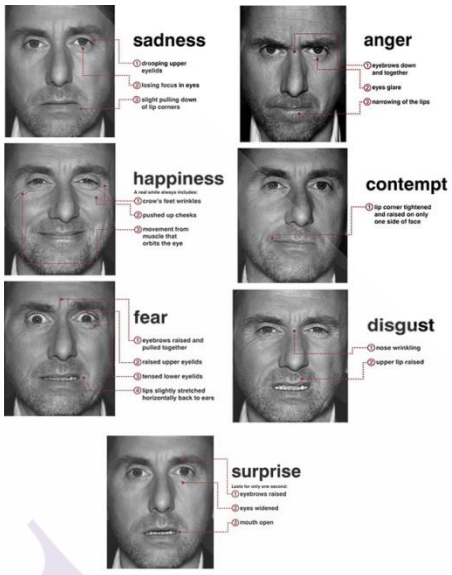
3. ความโกรธ (Anger) เวลาที่เรารู้สึกโกรธ หัวคิ้วจะย่นเข้าหากัน ทำให้เกิดริ้วรอยในแนวตั้งที่เห็นได้ชัดเจน ซึ่งเรียกกันว่า "คิ้วพันกัน" นอกจากนี้ยังอาจทำให้เปลือกตาด้านบนเบิกขึ้น ทำให้เห็นตาขาว และเกิดการแสดงออกที่เครียดเขม็ง เมื่อเปลือกตาล่างเกร็งเข้าหากัน ริมฝีปากเม้มแน่นและมองไม่เห็นริมฝีปากบนในส่วนที่เป็นสีชมพู

4. ความกลัว (Fear) เมื่อเรารู้สึกกลัว คิ้วของเราจะแบนเกือบราบ เกิดรอยย่นบนเส้นคิ้วเช่นเดียวกับความโกรธ ตาจะเบิกกว้าง เปลือกตาบนเปิด ทำให้มองเห็นตาขาว มุมปากฉีกออกไปด้านข้างเป็นเส้นตรงแน่น

5. ประหลาดใจ (Surprise) เมื่อเราประหลาดใจ เปลือกตาบนจะถูกยกขึ้น ทำให้เห็นตาขาวและอ้าปากหรือขากรรไกร

6. รังเกียจ (Disgust) เมื่อเรารู้สึกรังเกียจ เราจะย่นจมูกเหมือนเวลาได้กลิ่นเหม็น ทำให้เกิดรอยย่นแนวนอนที่บริเวณจมูกใกล้หว่างคิ้ว มีการแบะริมฝีปากบนร่วมด้วย

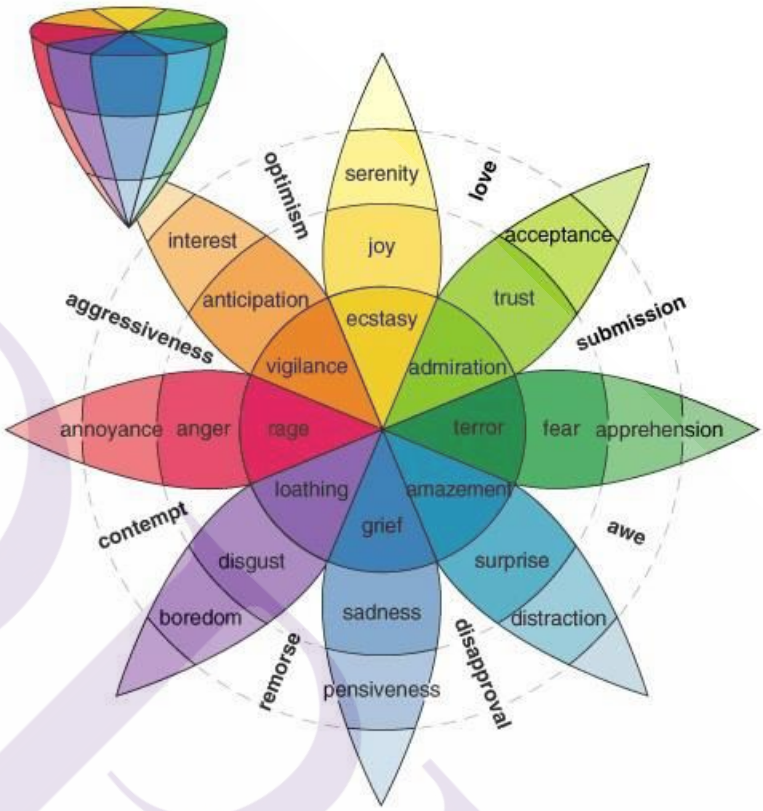
7. เกลียด (Contempt) เมื่อเรารู้สึกเกลียด มุมปากซ้ายจะฉีกออกด้านข้าง เกิดเป็นรอยบุ๋มและโดยมากจะกลอกตาร่วมด้วย ดังภาพที่ 2.1



ภาพที่ 2.1 การแสดงออกทางสีหน้า (Facial Expression) ของ Paul Ekman

ที่มา: Paul Ekman Facial Action Coding System, (1934)

Mulder, P. (2018) ได้ทำการศึกษาวิจัยเรื่องอารมณ์และเชื่อว่าอารมณ์มีพื้นฐานอยู่ 8 ชนิด ของ Robert คือ กลัว ประหลาดใจ เศร้า รังเกียจ โกรธ คาดหวัง รื่นเริง และยอมรับ อารมณ์ทั้ง 8 นี้ยังแปรเปลี่ยนไปตามระดับความความเข้มข้นของอารมณ์ โดยแต่ละอารมณ์ก็เกิดมาจากลำดับขั้นของความรู้สึกที่ใกล้กัน ตัวอย่างเช่น อารมณ์รัก (love) อยู่ระหว่าง กลีบสีเหลือง+เขียว นั่นก็คือ ความแจ่มใส+การยอมรับ (Serenity + Acceptance) ซึ่งกว่าจะถึงขั้นที่เรารู้สึกถึง Serenity ก็จะมีความรู้สึกที่ใกล้เคียงกันเกิดขึ้น ได้แก่ ecstasy ความปิติยินดี ถึงขีดสุด joy ความรู้สึกเป็นสุข รื่นเริง และกว่าจะมาถึงขั้น Acceptance นั้น ก็จะต้องเกิดความรู้สึกสองอย่างนี้ คือ admiration ได้รับความชื่นชม trust ไว้เนื้อเชื่อใจ และเมื่อสองกลีบนี้ (เหลือง+เขียว) รวมกันแล้ว จึงกลายเป็นอารมณ์รัก (Love) เกิดขึ้นได้ แต่ถ้าเกิดไปมีความรู้สึกแบบกลีบเขียว+เขียวเข้ม ก็จะไม่ใช่อารมณ์รัก แต่จะกลายเป็นอารมณ์ยอมเพราะกลัว (Submission) เป็นต้น ดังภาพที่ 2.2



ภาพที่ 2.2 แสดงระดับความเข้มของอารมณ์พื้นฐานของ Robert Plutchik

ที่มา: Mulder, P. (2018)

แพงค์เซปป์ (Jaak Panksepp, 1998) ได้เสนอแนวคิดในการจำแนกอารมณ์ที่แตกต่างจากพลูทชิค แพงค์เซปป์เชื่อว่า อารมณ์พื้นฐานมีอยู่ 4 ชนิดคือ คาดหวัง (Expectancy) เดือดดาล (Rage) ตื่นตระหนก (Panic) หวาดกลัว (Fear) อารมณ์พื้นฐานแต่ละชนิดเกิดขึ้นสัมพันธ์กับตำแหน่งในสมองที่ตั้งอยู่บริเวณไฮโปธาลามัส ซึ่งแต่ละจุดบนสมองจะสนองตอบต่ออารมณ์ต่างชนิดกัน โดยอาศัยการแปลข้อมูลที่ถูกกระตุ้นจากสิ่งแวดล้อมเป็นกระแสประสาทส่งไปยังประสาทมอร์เตอร์ จึงทำให้มนุษย์แปลอารมณ์ออกมาแตกต่างกัน เพื่อให้ง่ายต่อการทำความเข้าใจเรื่อง ความรู้สึกที่แท้จริงของตนเองและผู้อื่น จึงได้จำแนกอารมณ์ออกเป็น 2 กลุ่มใหญ่ๆ คือ

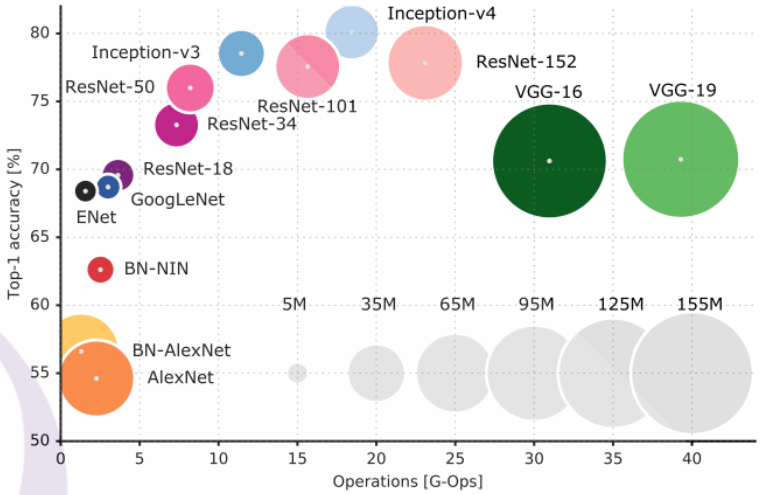
1. อารมณ์ที่ทำให้เกิดความพึงพอใจ (Pleasantness) คือ มีความสุข ต้องการให้เกิดขึ้น ต้องการยืดเหนี่ยวไว้เป็ นอารมณ์ทางบวก เช่น รื่นเริง ชื่นชม รัก ยอมรับ ฯลฯ เป็นต้น

2. อารมณ์ที่ทำให้เกิดความไม่พึงพอใจ (Unpleasantness) คือ มีความทุกข์ต้องการหลีกเลี่ยง ไม่ต้องการให้เกิดขึ้ นอีก เป็ นอารมณ์ทางลบ เช่น กลัว เศร้า เกลียด ขยะแขยง เดือดดาล ดูถูก อิจฉาริษยา ฯลฯ เป็นต้น

จึงสามารถสรุปได้ว่า การจัดประเภทอารมณ์ที่แสดงออกทางสีหน้าของมนุษย์ เป็นการจำแนกอารมณ์เป็นเรื่องยุ่งยาก เพราะคนเรามีสภาวะอารมณ์หลายอย่างเปลี่ยนแปลงไปตลอดเวลา และอารมณ์แต่ละชนิดก็ยังมีระดับความรุนแรงแตกต่างกันไป ยกตัวอย่างเช่น อารมณ์โกรธจะเริ่มจากความรู้สึกรำคาญใจ ไปจนถึงโกรธเกรี้ยวเดือดดาล หรืออารมณ์กลัวจะมีความรุนแรงตั้งแต่หวั่นกลัวไปจนถึงกลัวอย่างขนพองสยองเกล้า นอกจากนี้เรายังมีอารมณ์หลายๆ อารมณ์ในเวลาเดียวกันจนแยกไม่ออกว่าเป็นอารมณ์อะไรบ้าง บอกได้แต่เพียงว่า รู้สึกไม่สบายใจ เช่น ความวิตกกังวลซึ่งเป็นความรู้สึกที่คละเคล้าปนเปกันระหว่างความกลัว โกรธ เศร้า สำนึกผิด อับอาย เป็นต้น

2.2 สถาปัตยกรรมเครือข่ายประสาทเทียม (Neural Network Architectures)

Eugenio Culurciello (2017) เครือข่ายประสาทเชิงลึกและการเรียนรู้อย่างลึกซึ้ง (Deep Learning) เป็นขั้นตอนวิธีที่มีประสิทธิภาพและเป็นที่นิยม ซึ่งในปัจจุบันประสบความสำเร็จอย่างมากในการออกแบบสถาปัตยกรรมเครือข่ายประสาท จากการวิเคราะห์รูปแบบเครือข่ายประสาทเทียมลึกสำหรับการประยุกต์ใช้ในทางปฏิบัติ ได้มีการศึกษาโดย Alfredo Canziani, Adam Paszke และ Eugenio Culurciello (2016) ซึ่งได้ทำการศึกษาและเปรียบเทียบความถูกต้องเพื่อจัดอันดับ ของ Top accuracy และ Operation (G-Ops) ของแต่ละสถาปัตยกรรมเครือข่ายประสาทเทียมที่เป็นที่นิยม ดังภาพที่ 2.3



ภาพที่ 2.3 Reporting top-1 one-crop accuracy

ที่มา: https://towardsdatascience.com/neural-network-architectures-156e5bad51ba

2.2.1 โครงข่ายประสาทเทียมแบบ Multilayer Perceptron

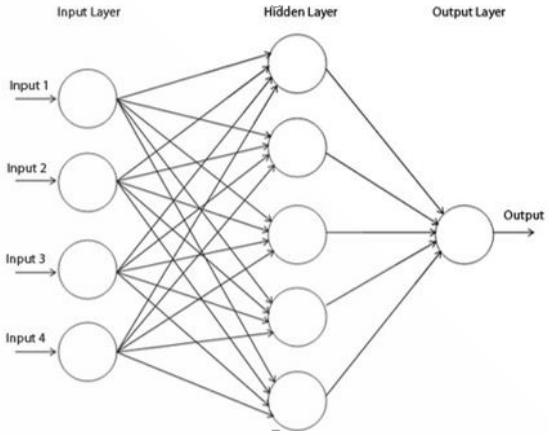
Syed Danish Ali and Rahul Ahuja (2016) โครงข่ายประสาทเทียมแบบ MLP เป็นรูปแบบหนึ่งของโครงข่ายประสาทเทียมที่มีโครงสร้างเป็นแบบหลายๆชั้น ใช้สำหรับงานที่มีความซับซ้อนได้ผลเป็นอย่างดี โดยมีกระบวนการฝึกฝนเป็นแบบมีผู้สอน (Supervise) และใช้ขั้นตอนการส่งค่าย้อนกลับ (Backpropagation) สำหรับการฝึกฝนกระบวนการส่งค่าย้อนกลับ ประกอบด้วย 2 ส่วนย่อย คือ การส่งผ่านไปข้างหน้า (Forward Pass) การส่งผ่านย้อนกลับ (Backward Pass) สำหรับการส่งผ่านไปข้างหน้า ข้อมูลจะผ่านเข้าโครงข่ายประสาทเทียมที่ชั้นข้อ มูลเข้า และจะส่งผ่าน จากอีกชั้นหนึ่งไปสู่อีกชั้นหนึ่งจนกระทั่งถึงชั้นข้อมูลออก ส่วนการส่งผ่านย้อนกลับค่าน้ำหนักการเชื่อมต่อจะถูกปรับเปลี่ยนให้สอดคล้องกับกฎการแก้ข้อผิดพลาด (Error-Correction) คือผลต่างของผลตอบที่แท้จริง (Actual Response) กับผลตอบเป้าหมาย (Target Response) เกิดเป็นสัญญาณผิดพลาด (Error Signal) ซึ่งสัญญาณผิดพลาดนี้จะถูกส่งย้อนกลับเข้าสู่โครงข่ายประสาทเทียมในทิศทางตรงกันข้ามกับการเชื่อมต่อและค่าน้ำหนักของการเชื่อมต่อจะถูกปรับจนกระทั่งผลตอบที่แท้จริงเข้าใกล้ผลตอบเป้าหมายสัญญาณที่มีโครงข่ายประสาทเทียมแบบ MLP มี 2 ประเภทคือ Function Signal และ Error Signal

2.2.1.1 Function Signal เป็นสัญญาณเข้าที่มาจากโหนดในชั้นก่อนหน้า และจะส่งผ่านไปข้างหน้าจากโหนดหนึ่งไปสู่อีกโหนดหนึ่ง

2.2.1.2 Error Signal เป็นสัญญาณย้อนกลับที่เกิดขึ้นที่โหนดในชั้นข้อมูลออกของโครงข่ายประสาทเทียม และถูกส่งผ่านย้อนกลับจากชั้นหนึ่งไปสู่อีกชั้นหนึ่ง

หลักการทำงานของ MLP คือในแต่ละชั้นของชั้นซ่อนตัว (Hidden Layer) จะมีฟังก์ชันสำหรับคำนวณเมื่อได้รับสัญญาณ (Output) จากโหนดในชั้นก่อนหน้านี้ เรียกว่า Activation Function โดยในแต่ละชั้นไม่จำเป็นต้องเป็นฟังก์ชันเดียวกันก็ได้ ชั้นซ่อนตัวนั้นมีหน้าที่สำคัญคือ จะพยายามแปลงข้อมูลที่เข้ามาในชั้น (Layer) นั้นๆให้สามารถแยกแยะความแตกต่างโดยใช้เส้นตรงเส้นเดียว (Linearly Separable) และก่อนที่ข้อมูลจะถูกส่งไปถึงชั้นข้อมูลออก (Output Layer) ในบางครั้งอาจจำเป็นต้องใช้ชั้นซ่อนตัวมากกว่า 1 ชั้นในการแปลงข้อมูลให้อยู่ในรูป Linearly Separable

ในการคำนวณหา Output ในปัญหาการจำแนกทำได้โดยการใส่ข้อมูล Input เข้าไปในโครงข่ายประสาทเทียมที่เราได้ทำการหาไว้แล้ว จากนั้นให้ทำการเปรียบเทียบค่าของ Output ใน Output Layer และให้ทำการเลือกค่าของ Output ที่มีค่าสูงกว่า (Neuron ที่มีค่าสูงกว่า) และทำการรับค่าของพยากรณ์ที่ตรงกับ Neuron ที่เลือก และให้นำค่าของ มาเปรียบเทียบกับค่าที่ยอมรับได้ หากค่าของ อยู่ในช่วงที่รับได้ (Error น้อยกว่า Error ที่เรากำหนด) ก็ให้ทำการรับข้อมูลชุดถัดไป แต่หากค่าของมากกว่าค่าที่ยอมรับได้ ให้ทำการปรับค่าน้ำหนักและ Biased ตามขั้นตอนที่ได้กล่าวไว้ข้างต้น เมื่อทำการปรับน้ำหนักเรียบร้อยแล้ว ให้ทำการรับข้อมูลชุดถัดไปและทำตามขั้นตอนซ้ำอีกรอบจนกระทั่งถึงข้อมูลชุดสุดท้าย และเมื่อทำข้อมูลชุดสุดท้ายเสร็จจะนับเป็น 1 รอบของการคำนวณ (1 Epoch) จากนั้นจะทำการหาค่าผิดพลาดรวมเฉลี่ย จากค่าเฉลี่ยของ ที่ได้เก็บค่าเอาไว้ เพื่อใช้ในการตรวจสอบว่าค่า โดยเฉลี่ยในการจำแนกนั้น มีค่าน้อยกว่าค่าผิดพลาดที่ยอมรับได้หรือไม่ ถ้าใช่แสดงว่าโครงข่ายประสาทเทียมที่สร้างขึ้นนั้นสามารถให้ผลลัพธ์ที่ถูกต้องของทุกๆข้อมูลแล้ว จึงทำการจบการเรียนรู้ได้ แต่ถ้าไม่ใช่ ให้กลับไปทำตามขั้นตอนแรก โดยเริ่มรับข้อมูลชุดที่ 1 ใหม่ ดังภาพที่ 2.4

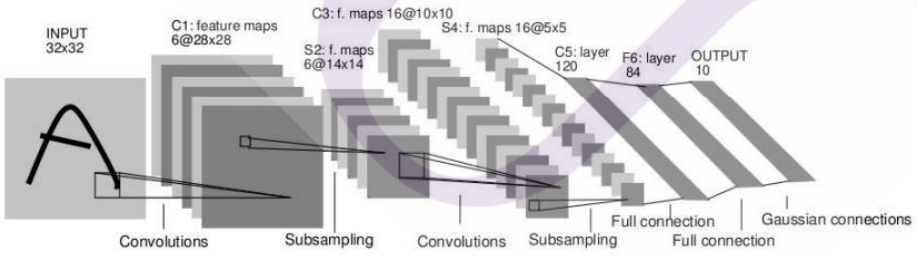


ภาพที่ 2.4 Multilayer Perceptron

ที่มา: Syed Danish Ali and Rahul Ahuja (2016)

2.2.2 ความรู้พื้นฐานเกี่ยวกับ Convolutional Neural Networks (CNNs)

ธนภัทร์ คุ้มสุภา (2559) เป็นนิวรอลเน็ตเวิร์กเชิงลึกรูปแบบหนึ่ง ซึ่งมีจุดเริ่มต้นมาจากงานวิจัยทางด้านการรู้จำภาพ โดยมักจะใช้ข้อมูลรับเข้าเป็นเมทริกซ์จากการแปลงมาจากรูปภาพ โครงสร้างของนิวรอลเน็ตเวิร์กคอนโวลูชัน แสดงดังภาพที่ 2.5 ซึ่งเน็ตเวิร์กทั้งหมดเกิดจากการนำชั้นหลายๆ ประเภทมาประกอบเข้าด้วยกันกังต่อไปนี้

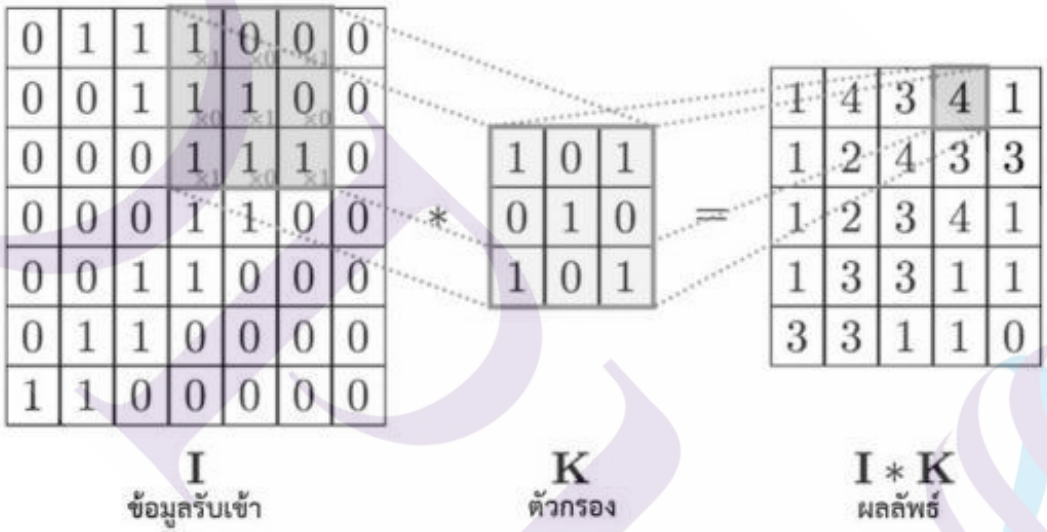


ภาพที่ 2.5 นิวรอลเน็ตเวิร์กคอนโวลูชัน

ที่มา: Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, (1998)

2.2.2.1 ชั้นคอนโวลูชัน (Convolutional Layer)

เป็นชั้นที่ทำการหาฟีเจอร์ของกลุ่มข้อมูลที่รับเข้ามาที่อยู่ใกล้ๆ กันโดยใช้วิธีการดอทเมทริกซ์กับตัวกรอง (Filter) โดยที่น้ำหนักของตัวกรอง (Filter) จะใช้ร่วมกันในทุกๆ การทำคอนโวลูชันของข้อมูลรับเข้ากำหนดให้ข้อมูลรับเข้าแทนด้วยเมทริกซ์ I และตัวกรอก (Filter) แทนด้วยเมทริกซ์ K ซึ่งมีขนาด h x w ผลลัพธ์ของการทำคอนโวลูชัน ดังภาพที่ 2.6



ภาพที่ 2.6 การทำคอนโวลูชัน (Convolutional)

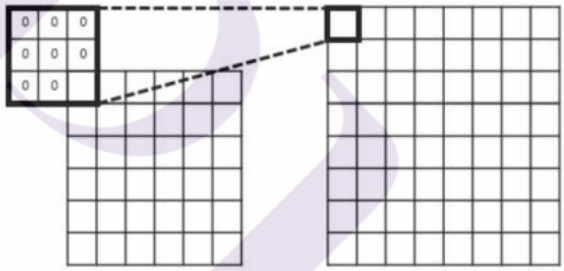
ที่มา: ธนภัทร์ คุ้มสุภา (2559)

ในชั้นคอนโวลูชัน มีองค์ประกอบที่ต้องพิจารณาดังต่อไปนี้

1) ขนาดของตัวกรอง (Filter Size) คือ ความกว้างและความสูงของตัวกรอง (Filter) ที่จะนำมาใช้ในการทำคอนโวลูชัน (ค่า h และ w) ซึ่งจากตัวอย่างภาที่ 2.5 ใช้ตัวกรอง (Filter Size) ที่มีขนาด 3x3 Filter

2) ชนิดของการทำคอนโวลูชัน (Convolution Type) คอนโวลูชันแบบแคบ (Narrow Convolution) โดยทั่วไปการทำคอนโวลูชันมักจะเป็นแบบแคบ ซึ่งในการทำคอนโวลูชัน ตัวกรอง (Filter) ที่นำมาทำการดอทเมตริส์นั้นจะไม่มีการกระทำเลยขอบของเมตริกซ์ข้อมูลรับเข้า ส่งผลให้ผลลัพธ์ที่ได้จากการทำคอนโวลูชันของข้อมูลรับเข้าที่มีขนาด NxN กับตัวกรอง (Filter) ที่มีขนาด MxM จะได้เมตริกซ์ขนาด (N-M+1)x(N-M+1) ตัวอย่างการทำคอนโวลูชันแบบแคบ ดังภาพที่ 2.5

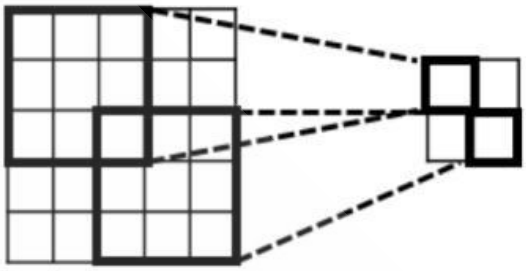
คอนโวลูชันแบบกว้าง (Wide Convolution) เป็นการทำคอนโวลูชันที่มีการกระทำเลยขอบของเมตริกซ์ข้อมูลรับเข้าออกไป โดยที่พื้นที่ส่วนที่เกิน ออกไปนั้นจะมีการแทนค่าของข้อมูล ณ ช่องนั้น ๆ ด้วย 0 ซึ่งเรียกว่า การเสริมเติม (Padding) ผลลัพธ์ที่ได้จากการ ทำคอนโวลูชันของข้อมูลรับเข้าที่มีขนาด NxN กับตัวกรองที่มีขนาด MxM จะได้เมตริกซ์ขนาด (N+M-1)x(N+M-1) ทั้งนี้การทำคอนโวลูชันแบบกว้างมีขึ้นเพื่อป้องกันการสูญเสียข้อมูลตรงบริเวณขอบของข้อมูลรับเข้า ตัวอย่างการทำคอนโวลูชันแบบกว้างดังภาพที่ 2.7



ภาพที่ 2.7 การทำคอนโวลูชันแบบกว้าง (Wide Convolution) และการเสริมเติม (Padding)

ที่มา: ธนภัทร์ คุ้มสุภา (2559)

ขนาดของการก้าวข้าม (Stride Size) คือจำนวนช่องของข้อมูลรับเข้า ที่จะทำการเลื่อนไปเมื่อทำการหาผลลัพธ์ของการคอนโวลูชันในแต่ละช่องโดยทั่วไปมักจะใช้ขนาดของการก้าวข้ามเป็น 1 ตัวอย่างการทำคอนโวลูชันที่มีขนาดของการก้าวข้ามเป็น 1 ดังภาพที่ 2.7 และภาที่ 2.8 จะแสดงลักษณะของการทำคอนโวลูชันที่มีขนาดของการก้าวข้ามเป็น 2

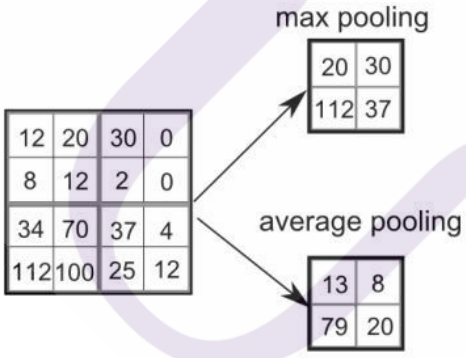


ภาพที่ 2.8 การทำคอนโวลูชันโดยมีข้อมูลรับเข้าขนาด 5x5 ตัวกรองขนาด 3x3 และมีขนาดของการก้าวข้ามเป็น 2

ที่มา: ธนภัทร์ คุ้มสุภา (2559)

2.2.2.2 ชั้นการรวม (Pooling Layer)

ทำหน้าที่ลดขนาดของข้อมูล เพื่อให้เหลือเฉพาะข้อมูลที่สำคัญๆ เท่านั้น ซึ่งมักจะนิยมนำมาต่อกับชั้นคอนโวลูชัน โดยทั่วไปนิยมใช้การเลือกข้อมูลที่มีค่ามากที่สุด (Max Pooling) หรือ ค่าเฉลี่ย (Average Pooling) มาจากแต่ละช่วงของเมตริกซ์เพื่อสร้างเป็นเมทริกซ์ที่มีขนาดเล็ก ดังภาพที่ 2.9

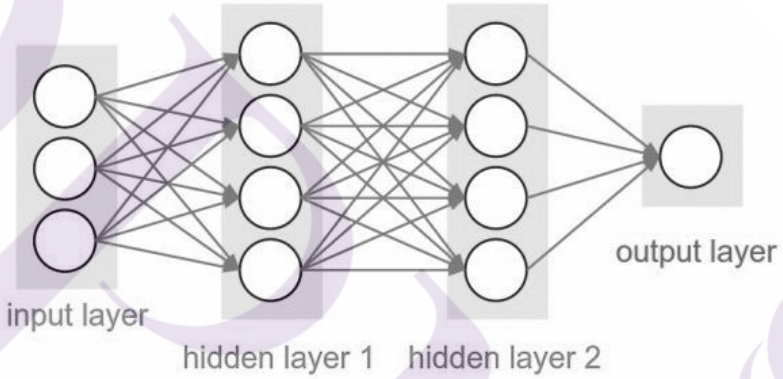


ภาพที่ 2.9 ตัวอย่างชั้นการรวมโดยค่าที่มากที่สุดและค่าเฉลี่ย

ที่มา: ธนภัทร์ คุ้มสุภา (2559)

2.2.2.3 ชั้นการเชื่อมโยงเต็มรูปแบบ (Fully Connected Layer)

หลังจากการประกอบกันของชั้นคอนโวลูชันและชั้นการรวมจำนวนหนึ่งแล้ว ในชั้นสุดท้ายของนิวรอลเน็ตเวิร์กคอนโวลูชันจะเป็นการเชื่อมโยงเต็มรูปแบบ นั่นคือ ในชั้นนี้จะประกอบด้วยชั้นย่อยๆ ที่มีเพอร์เซ็ปตรอนอยู่ จำนวนหนึ่ง โดยที่เพอร์เซ็ปตรอนแต่ละตัว จะมีเส้นเชื่อมกันเพอร์เซ็ปตรอน ทุกตัวในชั้นก่อนหน้าและเพอร์เซ็ฟตรอน ทุกตัวในชั้นถัดไป ทำให้สามารถทำการคำนวณการป้อนไปข้างหน้าและการแพร่กระจายย้อนกลับได้ด้วยวิธีการปกติได้ชั้นการเชื่อมโยงเต็มรูปแบบ ดังภาพที่ 2.10



ภาพที่ 2.10 ชั้นการเชื่อมโยงเต็มรูปแบบ (Fully Connected Layer)

ที่มา: ธนภัทร์ คุ้มสุภา (2559)

2.3 สถาปัตยกรรมนิวรอลเน็ตเวิร์กเชิงลึก CNNs Architecture

2.3.1 VGG เป็นสถาปัตยกรรมเครือข่ายประสาทเทียมรูปแบบหนึ่ง ซึ่ง VGG ถูกคิดค้นจาก Oxford เป็นคนแรกที่ใช้ตัวกรองขนาด 3 × 3 (filters) ในแต่ละชั้นของ convolutions และรวมกันเป็นลำดับของ convolutions ดังภาพที่ 2.11

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Table 2: **Number of parameters** (in millions).

Network	A,A-LRN	B	C	D	E
Number of parameters	133	133	134	138	144

ภาพที่ 2.11 Neural Network Architectures: VGG

ที่มา: https://towardsdatascience.com/neural-network-architectures-156e5bad51ba

เครือข่าย VGG นี้จะใช้ multiple convolutiona layers ขนาด 3x3 1 หลายแบบ เพื่อแสดงถึงคุณลักษณะที่ซับซ้อน สังเกตจากภาพที่ 2.9 บล็อก 3, 4, 5 ของ VGG-E: 256 × 256 และ 512 × 512 ตัวกรอง 3 × 3 filters ถูกใช้หลายครั้งในลำดับเพื่อแยกคุณลักษณะที่ซับซ้อนและการรวมกันของคุณสมบัติดังกล่าวนี้มีประสิทธิภาพ เช่น มีขนาดใหญ่ 512 × 512 classifiers กับ 3 ชั้น ซึ่งเป็น convolutional สิ่งนี้

เห็นได้ชัดว่ามีจำนวนพารามิเตอร์และมีความสามารถในการเรียนรู้อย่างมาก แต่การฝึกอบรมเครือข่ายเหล่านี้เป็นเรื่องยากและต้องแยกออกเป็นเครือข่ายขนาดเล็กที่มีการเพิ่มทีละชั้น ทั้งหมดนี้เนื่องจากไม่มีวิธีที่ดีในการจัดรูปแบบให้เป็นระเบียบหรือเพื่อจำกัดพื้นที่การค้นหาขนาดใหญ่ที่ได้รับการประชาสัมพันธ์โดยใช้พารามิเตอร์จำนวนมาก

สถาปัตยกรรมเครือข่ายประสาทเทียม VGG ใช้คุณลักษณะขนาดใหญ่ในหลายชั้น และต้องใช้ระยะเวลาในการทำงานน่อนข้างนาน

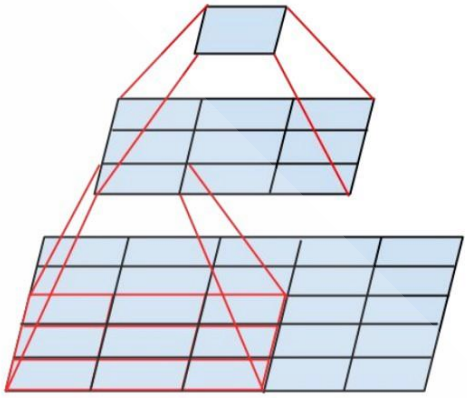
2.3.2 Inception V3, V2 จากการพัฒนาอย่างต่อเนื่องจากสถาปัตยกรรมเครือข่ายประสาทเทียม Inception ทำให้เกิด สถาปัตยกรรมเครือข่ายประสาทเทียม Inception V2 และ V3 ในปี 2015 ทำให้ Inception ได้รับการแนะนำให้เป็น Inception V2 จากการคำนวณโดยการใช้ Batch-normalization ในการคำนวณค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐาน ของคุณลักษณะทั้งหมดที่เอาต์พุตเลเยอร์หรือของชั้นถัดมา ในปี 2015 ได้เปิดตัว Inception Module ซึ่งเป็นสถาปัตยกรรมเครือข่ายประสาทเทียมรุ่นใหม่โดย GoogLeNet ได้มีการพัฒนาในการทำงานที่มีรายละเอียดมากขึ้นกว่าเดิม สามารถอธิบายได้ดังนี้

1) มีการเพิ่มประสิทธิภาพการไหลเวียนข้อมูลเข้าสู่เครือข่ายโดยการสร้างเครือข่ายที่มีความลึกและความกว้าง ก่อนที่จะรวบรวมข้อมูลคุณสมบัติ

2) เมื่อสถาปัตยกรรมเครือข่ายประสาทเทียมมีความลึกเพิ่มขึ้น จำนวนของคุณสมบัติหรือความกว้างของเลเยอร์จะเพิ่มขึ้นอย่างเป็นระบบ

3) ใช้การเพิ่มความกว้างในแต่ละเลเยอร์เพื่อเพิ่มการรวมกันของคุณสมบัติก่อน ในเลเยอร์ถัดไป

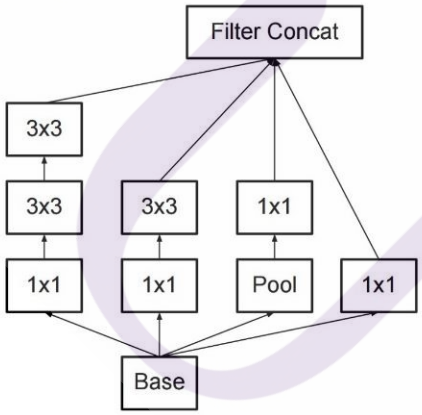
4) มีการใช้ 3x3 convolution เท่านั้น เมื่อเป็นไปได้ให้ตัวกรอง 5x5 และ 7x7 สามารถย่อยสลายได้ด้วย 3x3 หลายๆ ตัว ดังภาพที่ 2.12



ภาพที่ 2.12 That filter of 5x5 and 7x7 can be decomposed with multiple 3x3

ที่มา: Eugenio Culurciello, (2017)

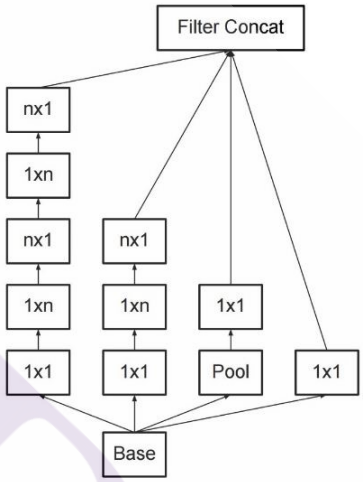
จึงทำให้สถาปัตยกรรมเครือข่ายประสาทเทียม Inception V3 ดังภาพ 2.12



ภาพที่ 2.13 the new inception module

ที่มา: Eugenio Culurciello, (2017)

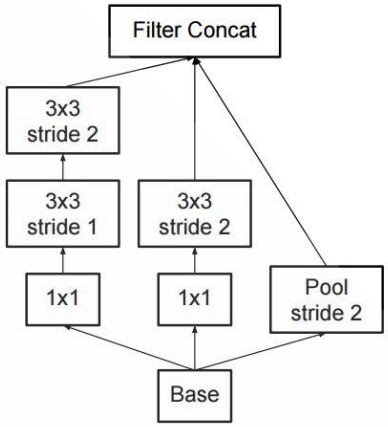
5) ทำให้ตัวกรอง (filter) ยังสามารถย่อยสลายโดย convolutions แบนเป็นโมดูลที่ซับซ้อนมากขึ้น ดังภาพที่ 2.14



ภาพที่ 2.14 Decomposed by flattened convolutions

ที่มา: Eugenio Culurciello, (2017)

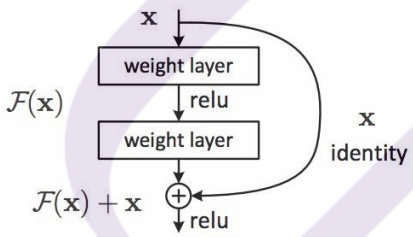
6) โมดูล Inception ยังมีความสามารถในการลดขนาดของข้อมูลโดยการ pooling ในขณะการคำนวณ ซึ่งเป็นพื้นฐานเหมือนกันกับการดำเนินการ convolution แบบคู่ขนาน ในการทำ simple pooling layer ดังภาพที่ 2.15



ภาพที่ 2.15 Decrease the size of the data

ที่มา: Eugenio Culurciello, (2017)

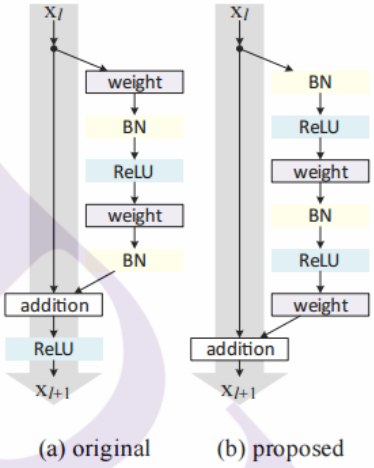
2.3.3 ResNet ถูกคิดค้นขึ้นมาพร้อมๆ กันกับ Inception v3 โดยมีการ feed ข้อมูลหรือผลลัพธ์ ของทั้งสองชั้น convolutional อย่างต่อเนื่อง และยังหลีกเลี่ยงการป้อนข้อมูลไปยังชั้นถัดไป ดังภาพที่ 2.16



ภาพที่ 2.16 Feed the output of two successive convolutional layer

ที่มา: Eugenio Culurciello, (2017)

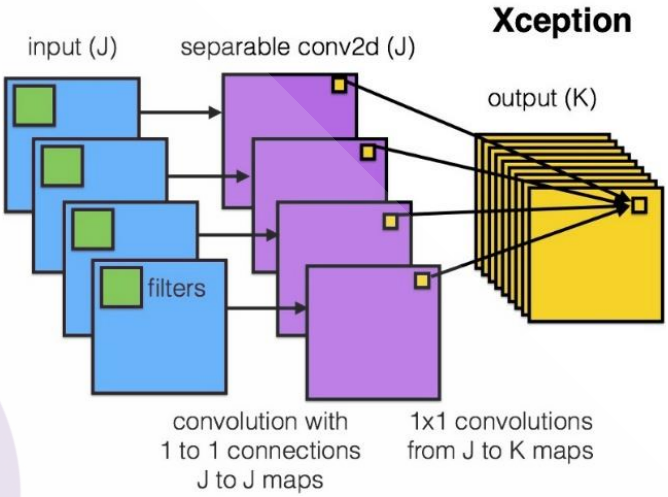
ResNet เป็นการนำ residual modules มาต่อกัน และใช้ stochastic descent gradient มาเทรนตัว input จะถูก preprocess โดยการแบ่งเป็น patch เล็กๆ ก่อนถูกนำเข้ามาในโมเดล ResNet เป็นหนึ่งในสถาปัตยกรรม monster ซึ่งกำหนดความลึกของสถาปัตยกรรมการเรียนรู้ที่ลึกซึ้งได้อย่างแท้จริง Residual Networks (ResNet in short) ประกอบด้วยโมดูลที่เหลือจำนวนมากตามมาซึ่งเป็นโครงสร้างพื้นฐานของสถาปัตยกรรม ResNet ดังภาพที่ 2.17



ภาพที่ 2.17 Residual Networks (ResNet in short)

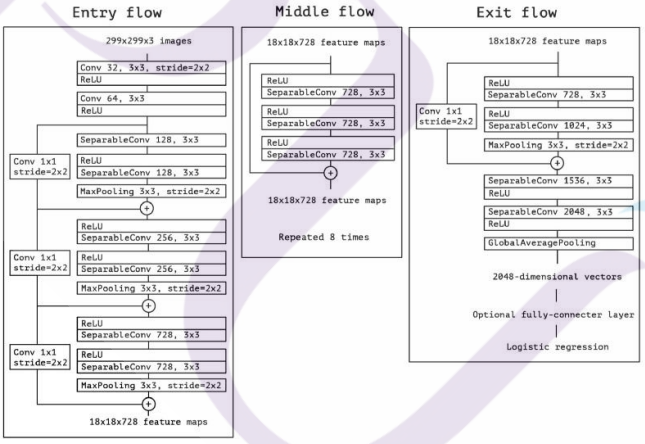
ที่มา: Eugenio Culurciello, (2017)

2.3.4 Xception François Chollet (2016) ได้ออกแบบโมดูล Xception ขึ้นมา แล้วพบว่ามีความสามารถช่วยปรับปรุงโมดูลและสถาปัตยกรรมการเริ่มต้นด้วยสถาปัตยกรรมที่เรียบง่ายและดีมากขึ้นซึ่งมีประสิทธิภาพเท่ากับ ResNet และ Inception V4 แต่มีการทำงานที่น้อยลงมากกว่า ดังภาพที่ 2.18 และ 2.19



ภาพที่ 2.18 The Xception module

ที่มา: Eugenio Culurciello, (2017)



ภาพที่ 2.19 The Xception architecture

ที่มา: Eugenio Culurciello, (2017)

สถาปัตยกรรมเครือข่ายประสาทเทียม Xception มีจำนวนทั้งหมด 36 convolutional ขั้นตอนการทำงานใกล้เคียงกับ ResNet-34 แต่รูปแบบและรหัสเป็นง่ายกว่า ResNet และเข้าใจได้มากขึ้นกว่า Inception V4 จึงทำให้สถาปัตยกรรมเครือข่ายประสาทเทียม Xception เป็นที่น่าสนใจมากขึ้น

2.3.5 E-Net

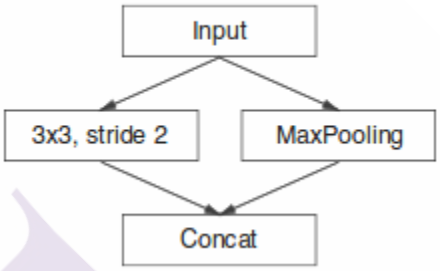
Abhishek Chaurasia, Sangpil Kim and Eugenio Culurciello (2016) เป็นเครือข่ายประสาทที่มีประสิทธิภาพอีกรูปแบบหนึ่ง ซึ่ง E-Net นี้มีจุดมุ่งหมายเพื่อให้สามารถใช้งานได้บนอุปกรณ์เคลื่อนที่ที่ใช้พลังงานต่ำและสามารถประมวลผลได้ที่มีความแม่นยำ สถาปัตยกรรมทั้งหมดของ E-Net ส่วนใหญ่จะมีลักษณะคล้ายกับ ResNets ที่มีโครงสร้างที่แยกออกเป็นหลายสาขาที่แยกออก แต่ยังผสานกลับผ่านการเพิ่มคุณลักษณะให้ชัดเจนมากขึ้นได้ เช่นเดียวกับโมเดลต้นแบบของ ResNet ตัวอย่างโครงสร้างดังภาพที่ 2.20

Name	Type	Output size
initial		$16 \times 256 \times 256$
bottleneck1.0	downsampling	$64 \times 128 \times 128$
$4\times$ bottleneck1.x		$64 \times 128 \times 128$
bottleneck2.0	downsampling	$128 \times 64 \times 64$
bottleneck2.1		$128 \times 64 \times 64$
bottleneck2.2	dilated 2	$128 \times 64 \times 64$
bottleneck2.3	asymmetric 5	$128 \times 64 \times 64$
bottleneck2.4	dilated 4	$128 \times 64 \times 64$
bottleneck2.5		$128 \times 64 \times 64$
bottleneck2.6	dilated 8	$128 \times 64 \times 64$
bottleneck2.7	asymmetric 5	$128 \times 64 \times 64$
bottleneck2.8	dilated 16	$128 \times 64 \times 64$
Repeat section 2, without bottleneck2.0		
bottleneck4.0	upsampling	$64 \times 128 \times 128$
bottleneck4.1		$64 \times 128 \times 128$
bottleneck4.2		$64 \times 128 \times 128$
bottleneck5.0	upsampling	$16 \times 256 \times 256$
bottleneck5.1		$16 \times 256 \times 256$
fullconv		$C \times 512 \times 512$

ภาพที่ 2.20 Architecture of ENet.

ที่มา: ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation (2016)

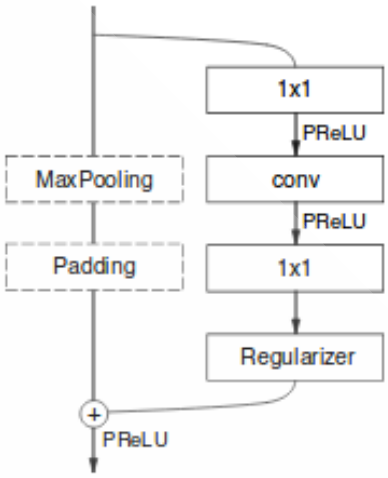
ส่วนเริ่มต้นของ E-Net (Initial block) คือ การนำข้อมูลหรือภาพ เข้าสู่โมเดลมีความละเอียด 512 x 512 ซึ่งส่งผลให้มีขนาดการแสดงผลของบล็อกเริ่มต้นที่ 16 x 256 x 256 หลังจากที่มีการรวมตัวของ convolution (13 filters) และ MaxPooling (2 x 2 without overlap) การสร้างภาพของบล็อกเริ่มต้นดังภาพที่ 2.21



ภาพที่ 2.21 Initial block of ENet.

ที่มา: ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation (2016)

ในส่วนต่อมาเป็นโมดูล Bottleneck มีโครงสร้างเหมือนกันภาพที่ 2.22 แต่ละสาขาประกอบด้วย 3 ชั้น การทำ convolutional ครั้งแรกเพื่อช่วยลดขนาดของข้อมูลลงด้วยขนาด 1 x 1 ซึ่งทำให้ระหว่าง convolution เหล่านี้มักเกิดการขยายตัว การทำ Batch normalization และ PReLU จะถูกวางไว้ระหว่าง convolution ทั้งหมด จากนั้นใช้ Dropout เชิงพื้นที่ ดังภาพที่ 2.22



ภาพที่ 2.22 Bottleneck module of ENet.

ที่มา: ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation (2016)

2.3.6 MobileNet

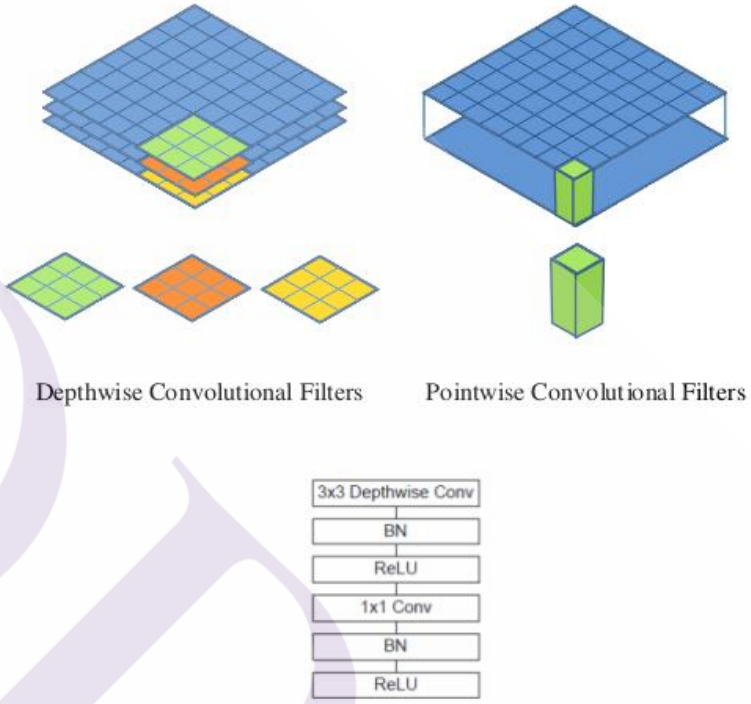
MobileNet เป็นโมเดลที่ถูกออกแบบมาสำหรับมือถือโดย Google โดยมาพร้อมกับ Weight ของโมเดลด้วยข้อมูลจากฐานข้อมูล ImageNet ที่ภายในรวบรวมภาพจำนวนหลายล้าน ซึ่งโมเดลนี้รองรับงานได้หลากหลาย MobileNet มีจุดประสงค์ในการออกแบบให้ใช้งานได้ โดยไม่ใช้ทรัพยากรมากนัก แต่ก็แลกกับประสิทธิภาพที่อาจจะน้อยลงถ้าเทียบกับ model ตัวอื่นๆ (Inception, Resnet) แต่ประสิทธิภาพก็อยู่ในเกณฑ์ที่ยอมพอรับได้

ตารางที่ 2.1 เปรียบเทียบ Documentation for individual models

The top-1 and top-5 accuracy refers to the model's performance on the ImageNet Validation dataset.

Model	Size	Top-1 Accuracy	Top-5 Accuracy	Parameters	Depth
Xception	88 MB	0.790	0.945	22,910,480	126
VGG16	528 MB	0.713	0.901	138,357,544	23
VGG19	549 MB	0.713	0.900	143,667,240	26
ResNet50	98 MB	0.749	0.921	25,636,712	-
ResNet101	171 MB	0.764	0.928	44,707,176	-
ResNet152	232 MB	0.766	0.931	60,419,944	-
ResNet50V2	98 MB	0.760	0.930	25,613,800	-
ResNet101V2	171 MB	0.772	0.938	44,675,560	-
ResNet152V2	232 MB	0.780	0.942	60,380,648	-
InceptionV3	92 MB	0.779	0.937	23,851,784	159
InceptionResNetV2	215 MB	0.803	0.953	55,873,736	572
MobileNet	16 MB	0.704	0.895	4,253,864	88
MobileNetV2	14 MB	0.713	0.901	3,538,984	88
DenseNet121	33 MB	0.750	0.923	8,062,504	121
DenseNet169	57 MB	0.762	0.932	14,307,880	169
DenseNet201	80 MB	0.773	0.936	20,242,984	201
NASNetMobile	23 MB	0.744	0.919	5,326,716	-

จากตารางที่ 2.1 เปรียบเทียบ Documentation for individual models พบว่า MobileNet ทำคะแนน Top-1, Top-5 ได้ 0.665 กับ 0.871 ตามลำดับ ซึ่งน้อยกว่าโมเดลตัวอื่นๆแต่ก็ถือว่าทำงานได้ดีพอสมควร



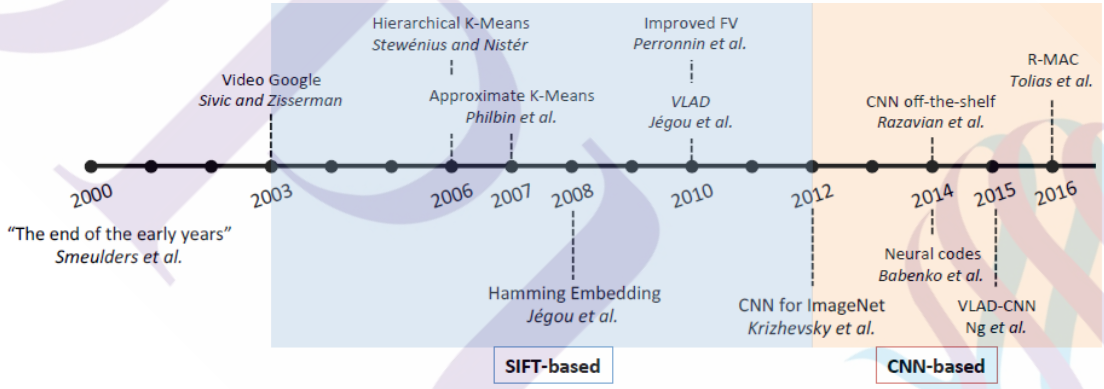
ภาพที่ 2.23 โครงสร้าง Model

ที่มา: https://github.com/Zehaos/MobileNet

2.4. ศึกษาเทคนิคการวิเคราะห์ใบหน้าแสดงอารมณ์

Liang Zheng, Yi Yang, and Qi Tian (2015). ได้ศึกษา SIFT Meets CNN: A Decade Survey of Instance Retrieval ได้เริ่มหลังจากการสำรวจเทคนิค ในปี ค.ศ. 2000 โดย Smeulders ต่อมาในปี ค.ศ. 2003 ได้มีการนำเสนอ Video Google โดย Sivic และ Zisserman ได้เป็นจุดเริ่มต้น ต่อมาในปี ค.ศ. 2006 ได้มีการศึกษาการทำงานแบบ hierarchical k-means โดย Philbin คือ การทำงานแบบการเรียงลำดับชั้นในการจัดกลุ่ม k-means โดยกำหนด k เป็นจำนวนกลุ่มที่สนใจ ต่อมาในปี ค.ศ. 2007 ได้มีการ

นำเสนอ approximate k-means โดย Stewénius และ Nistér คือ การแบ่งกลุ่มโดยการประมาณจำนวน k เพื่อให้ได้จำนวนกลุ่มที่สนใจ ต่อมาในปี ค.ศ. 2008 ได้มีการนำเสนอวิธีการ Hamming Embedding โดย Jégou คือ การแบ่งกลุ่มตามความคล้ายคลึงกัน ต่อมาในปี ค.ศ. 2010 ได้มีการนำเสนอเรื่อง improved FV การแสดงภาพขนาดกะทัดรัดสำหรับการดึงข้อมูล นำเสนอโดย Perronnin จึงได้เกิดเทคนิค SIFT-based (Scale Invariant Feature Transform) เป็นการวิเคราะห์รูปภาพเพื่อหาจุดเด่นของภาพที่ได้รับเข้ามาเปรียบเทียบ ซึ่งในช่วงเวลานี้ได้มีเทคนิคใหม่ค่อยๆ เกิดขึ้นคือ เทคนิค CNN-based (Convolutional Neural Networks) หลังจากการสำรวจในปี ค.ศ. 2012 โดย Krizhevsky เป็นครั้งแรก ต่อมาในปี ค.ศ. 2014 เริ่มมีการนำเสนอเทคนิควิธีการ hybrid ในองค์ประกอบของ CNN จากรูปภาพ โดย Razavian ในปีเดียวกัน Babenko ได้เป็นคนแรกในการปรับปรุงเทคนิค CNN ในการใช้กับข้อมูลที่มีอยู่ทั่วไป ดังภาพที่ 2.24



ภาพที่ 2.24 Milestones of instance retrieval

ที่มา: SIFT Meets CNN: A Decade Survey of Instance Retrieval

สามารถสรุปได้ว่า ในระยะเวลาตั้งแต่ปี ค.ศ. 2000-2016 สามารถแบ่งเทคนิควิธีการจากการทำสำรวจของ Liang Zheng, Yi Yang, and Qi Tian (2015) ได้เป็น 2 ส่วนคือ SIFT-based (Scale Invariant Feature Transform) และ CNN-based (Convolutional Neural Networks)

1 SIFT-based (Scale Invariant Feature Transform)

Lowe, David G. (1999) อธิบายไว้ว่า SIFT ที่ซึ่งย่อมาจาก Scale Invariant Feature Transform คือการเอาจุดเด่นในรูป ที่ขึ้นอยู่กับสเกลการกำหนดทิศทางตำแหน่ง มุมการมองแสงสว่างเงา มาทำให้จุดเด่นนั้นไม่ต้องขึ้นอยู่กับสเกลการกำหนดทิศทาง ฯลฯ ซึ่งจะทำให้สามารถนำมาใช้ในการเปรียบเทียบจุดเด่นในรูปอื่นๆ ได้ง่ายและถูกต้องแม่นยำมากยิ่งขึ้น โดยเป็นจุดเด่นที่มักนิยมใช้กันเป็นอย่างมากในการทำการระบุตำแหน่งที่ตั้ง

วิทูลย์ ดอนพรทัน และคณะ (2553) อธิบายไว้ว่า การตรวจหาลักษณะเด่นแบบซิฟท์ (SIFT) เป็นอัลกอริทึมที่นำมาใช้ในการวิเคราะห์เพื่อหาจุดเด่นของภาพ โดยไม่ขึ้นกับขนาดหรือทิศทางของวัตถุที่อยู่ในภาพซึ่ง สามารถนำมาประยุกต์ใช้ในเรื่องการรู้จำวัตถุ (object recognition) การทำงานของซิฟท์นั้นจะเริ่มจากการแปลงข้อมูลภาพเพื่อหาจุดที่มีลักษณะเด่น (keypoint) เพื่อใช้เป็นตัวบอกจุดลักษณะสำคัญ (keypoint descriptor) ที่มีลักษณะเด่นที่มากที่สุด เมื่อต้องการเทียบภาพวัตถุกับภาพที่ต้องการจำแนก ว่ามีวัตถุอยู่ในภาพดังกล่าวหรือไม่ ก็จะทำการหาจุดลักษณะเด่นและตัวบอกจุดลักษณะสำคัญ แล้วนำมาหาระยะทางแบบยุคลิด (Euclidean distance) ระหว่างตัวบอกจุดลักษณะสำคัญด้วยกัน ถ้าจุดที่อยู่ใกล้ที่สุด อยู่ใกล้กว่าระยะที่กำหนดไว้ก็จะถือว่าจุดลักษณะเด่นของทั้ง 2 ภาพนั้นเป็นจุดเดียวกัน วิธีการวิเคราะห์เพื่อหาจุดเด่นของภาพนั้นมีอยู่ 4 ขั้นตอนดังนี้

1) การหาปริภูมิค่าในมิติขนาดและระยะทาง (Scalespace extreme detection)

2) การกำหนดตำแหน่งจุดสนใจ (Keypoint localization)

3) การกำหนดทิศทางของจุดสนใจ (Orientation assignment)

4) การสร้างคำอธิบายลักษณะเด่นของภาพ (Keypoint descriptor)

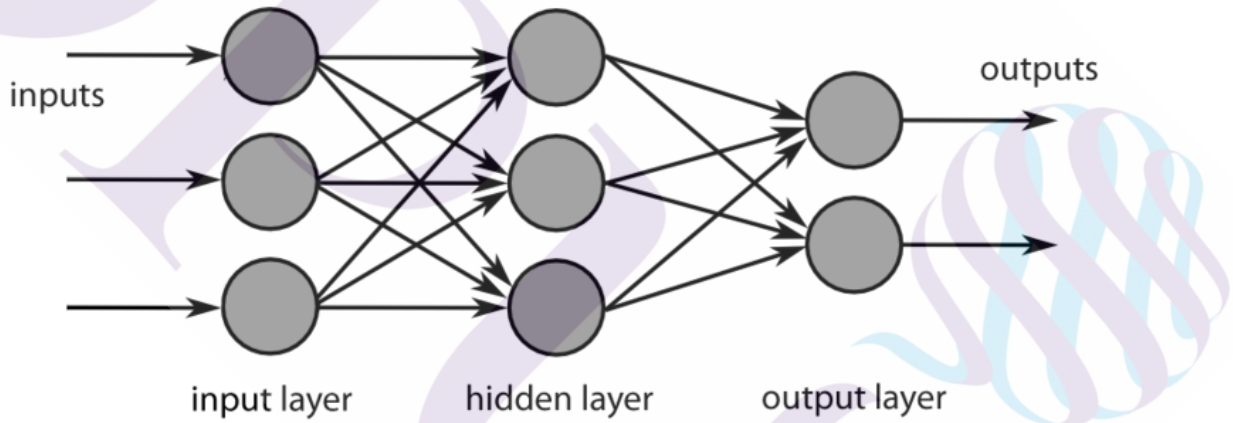
2 CNN-based (Convolutional Neural Networks)

ฐิติรัตน์ ศิริบวรรัตนกุล (2559) Convolutional Neural Network (CNN) เทคนิคมาตรฐานของ Deep Learning ที่นิยมใช้กันในปัจจุบัน ซึ่ง Deep Learning คือ รูปแบบหนึ่งของ Machine Learning ซึ่งจะว่าไปแล้วทั้งสองก็ไม่ใช่เรื่องใหม่อะไรถูกนำเสนอมาแล้วหลายสิบปี ไอเดียตั้งต้นของ Deep Learning มาจากแบบจำลอง Machine Learning ชนิดโครงข่ายประสาทเทียม (Artificial Neural Network) ที่เลียนแบบการทำงานของโครงข่ายเซลล์สมองของคนเรา ซึ่งแบบจำลองที่ว่านี้ได้รับการ

พิสูจน์ทางทฤษฎีมาในอดีตว่า ในกรณีที่โครงข่ายภายในลึก (Deep) และมีจำนวนโหนดหรือเซลล์มากพอ ระดับความซับซ้อนของแบบจำลองจะเพียงพอสำหรับแก้ปัญหาที่เราโยนเข้าไปได้ทุกชนิด

หลักการโดยทั่วไปของการเรียนรู้เชิงลึก Convolutional Neural Network (CNN) คือ การมีหน่วยประมวลผลหลายๆ ชั้น ข้อมูลขาเข้าในแต่ละชั้นได้มาจากปฏิสัมพันธ์กับชั้นอื่นๆ ทั้งนี้ การเรียนรู้เชิงลึกพยายามหาความสัมพันธ์ที่ล้ำลึกมากขึ้น นั่นคือ เมื่อมีจำนวนของชั้นและหน่วยประมวลผลที่อยู่ในชั้นมากขึ้น ข้อมูลในชั้นสูงๆก็จะยิ่งล้ำลึกซับซ้อนมากขึ้น โดยทั่วไป Neural Networks ประกอบไปด้วย Layer ทั้งหมด 3 Layer คือ

Input layer เป็น Layer รับข้อมูล hidden layer เป็น Layer ส่วนนี้เป็นส่วนเรียนรู้ของ Neural Networks (DEEP NN) output layer เป็น Layer รวมผลลัพธ์ที่ได้จาก hidden layer และแปลงข้อมูลเพื่อส่งออกข้อมูล ดังภาพที่ 2.25



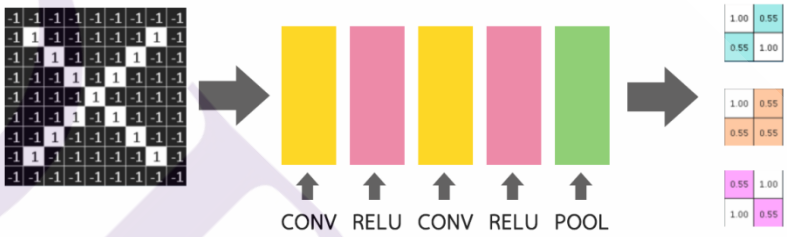
ภาพที่ 2.25 Deep Learning

ที่มา: MultiLayerNeuralNetwork

ในการฝึกฝนเครือข่ายประสาทเทียม Convolutional Neural Network นั้นเมื่อทำงานกับข้อมูลจำนวนมากและสถาปัตยกรรมเครือข่ายที่ซับซ้อน จำเป็นต้องใช้ Graphic Processing Unit (GPU) เพื่อช่วยเร่งความเร็วในการประมวลผลเพื่อฝึกแบบจำลองนั้นๆ

Core Layers ประกอบด้วย Dense คือ โครงสร้างแบบ fully connected ระหว่างชั้นที่ติดกัน โดยเราเลือกใช้ activation function และ Flatten ทำหน้าที่แปลงข้อมูลจากภาพหลาย channel ให้เป็นเวคเตอร์ ที่เราสามารถส่งต่อให้กับ Layers มาตรฐานต่อได้

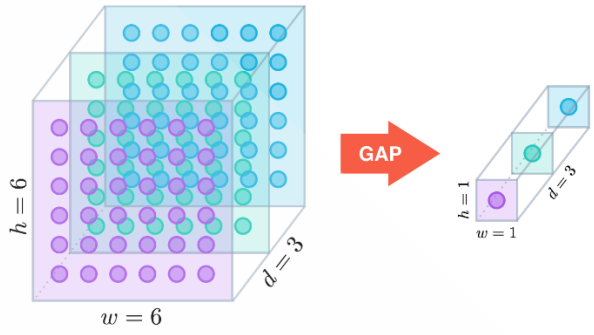
Pooling Layers ในส่วนของขั้นตอนการ Pooling นั้น จะทำการเลือกค่าที่มากสุด จากขนาดเมทริกซ์ ที่เลือกไว้ออกมา เพื่อปรับขนาดให้รูปภาพเล็กลง และได้คุณลักษณะที่สำคัญของข้อมูลออกมา ดังภาพที่ 2.26



ภาพที่ 2.26 กระบวนการทำงานของ Convolutional Neural Networks

ที่มา: MultiLayerNeuralNetwork

GlobalAveragePooling2D เป็นการลดจำนวนพารามิเตอร์ทั้งหมดลงในแบบจำลอง มีลักษณะคล้ายคลึงกับ maxing pooling layer เพื่อใช้ลดมิติข้อมูลเชิงพื้นที่ของเมตริกซ์ 3 มิติ อย่างไรก็ตาม maxing pooling layer มีการลดขนาดมากขึ้นโดยที่เมตริกซ์ที่มีขนาด h × w × d จะลดขนาดลงเป็นขนาด 1 × 1 × d เท่านั้น maxing pooling layer จะลดขนาดแผนที่แต่ละตัวลงเหลือเพียงจำนวนเดียวโดยใช้ค่าเฉลี่ยของค่าทั้งหมดของ h w ดังภาพที่ 2.27



ภาพที่ 2.27 global-average-pooling-layers-for-object-localization

ที่มา: https://alexisbcook.github.io/2017/global-average-pooling-layers-for-object-localization/

Convolution Layer : CONV2D ในขั้นตอนนี้จะทำการปรับฟิลเตอร์ให้ภาพ ด้วยฟิลเตอร์หลายแบบ เพื่อให้ได้ภาพในหลายๆคุณลักษณะ โดยคำนวณแต่ละค่าในพิกเซลของ feature ของภาพที่สอดคล้องกัน คูณแต่ละพิกเซลของ feature กับ filter โดยเป็นค่าของพิกเซลที่สอดคล้องกันในภาพ และบวกทุกคำตอบเข้าด้วยกันและหารด้วยจำนวนรวมของพิกเซลใน feature หากทั้งสองพิกเซลมีสีขาว (จะมีค่าเป็น 1)

Activations Rectified Linear Unit : RELU นำ output จาก ขั้นตอนการ Convolution ก่อนหน้ามาคำนวณจากนั้นผลลัพธ์ที่ได้คือค่าของน้ำหนักใหม่ สามารถแก้ปัญหา gradient vanishing ได้ เพราะค่า derivative ของมันคือ 1 ซึ่งทาให้ค่าเกรเดียนต์ที่ถูกส่งกลับมานั้นไม่โดนลดขนาดลง แต่ทั้งนี้ค่า derivative อีกค่าของมันคือ 0 ดังนั้นมันก็มีโอกาสที่จะโยนเกรเดียนต์นั้นทิ้งไปหมดก็ได้

Normalization Layers ที่ใช้ Batch Normalization (BN) ในที่นี้จะต้องรู้ว่าก่อนนำเอา feature vector มาวิเคราะห์ เราควรทำการ normalize ก่อน วิธีโดยการลบค่าเฉลี่ยให้มันมีการกระจายรอบ 0 และทำการ normalize ให้มีการกระจายตัวมาตรฐาน เช่นให้มีค่าเบี่ยงเบนมาตรฐานเป็น 1 เป็นต้น

Back propagation ในทุก layers ของ CNNs ที่ได้รับการฝึกฝนโดยใช้อัลกอริทึม back progagation สำหรับ error propagation และ weight adaptation ในการเชื่อมต่อแบบสลับซับซ้อนและ subsampling เลเยอร์ จากนั้นทำตามขั้นตอนมาตรฐาน ในชั้นรวมสูงสุด, สัญญาณผิดพลาดจะ

แพร่กระจายไปยังตำแหน่งที่ arg max ดังนั้นแผนที่ข้อผิดพลาดในชั้นรวมสูงสุดจะเบาบาง ซึ่งทำให้มีการสะสมสัญญาณผิดพลาดหลายชุดไว้ในเครื่องเดียว

เพื่อเพิ่มความเร็วในการ training ได้เลือกใช้ implemented แบบ CNN โดยใช้การประมวลผมจาก GPU โดยใช้ CUDA ของ NVIDIA ในการประมวลผล การดำเนินการแบบ Convolution ของ Alex Krizhevsky อธิบายไว้ว่า จะเป็นการเร่งให้การทำงานได้เร็วขึ้นด้วย CUV library สำหรับการเรียนรู้แบบ mini-batch ซึ่งในการประมวลผลแบบขนาน ทำให้สามารถเพิ่มความเร็วของคำสั่งได้ถึง 2 เท่า เมื่อเทียบกับการใช้ CPU ในการประมวลผล

Separable Convolution จากการเกิดปัญหาเกี่ยวกับชั้นของ convolution แบบดั้งเดิม ที่มีพารามิเตอร์จำนวนมากเกินไป เช่น ชั้น convolution 3x3 มีพารามิเตอร์ 9 ตัว และจำนวนนี้เพิ่มขึ้นทุกๆ 2 เท่าเมื่อเพิ่มขนาด kernel เพราะถ้ามีจำนวนพารามิเตอร์จำนวนมากจะทำให้ใช้เวลานานในการเรียนรู้พารามิเตอร์เหล่านี้ขณะฝึกอบรม ซึ่งในการลดจำนวนพารามิเตอร์และทำให้มีประสิทธิภาพนั้น ในกระบวนการ Separable Convolution จะทำโดยการคอมโพสิต (computational) เพื่อแยกพารามิเตอร์สามารถทำให้ช่วยลดจำนวนพารามิเตอร์ลงได้

วิธี Separable Convolution เป็นการคูณเวกเตอร์ที่ทำให้เราได้ผลลัพธ์ที่แน่นอน เราสามารถได้ผลลัพธ์ที่เหมือนกันโดยการคูณด้วยสองเวกเตอร์ที่เล็กๆ เช่น

ถ้าหาก convolution 3x3 จะมีพารามิเตอร์เท่ากับ 9 พารามิเตอร์ ซึ่งวิธีการของ Separable Convolution คือ การแยกคำนวณ โดยการหมุนวน จะได้เป็น convolution 1x3 และตามด้วย convolution 3x1 ซึ่งสามารถลดจำนวนพารามิเตอร์ลงเหลือ 6 พารามิเตอร์ เพื่อช่วยลดต้นทุนด้านการคำนวณมากขึ้น

Separable Convolution ได้ถูกนำมาใช้อย่างกว้างขวาง เช่นใน Xception CNN ได้รับการออกแบบโดย F Chollet ซึ่งเป็นผู้เขียน Keras Deep Learning Library ในการทำงานของ Xception ได้รับการออกแบบเพื่อให้ทราบว่าสถาปัตยกรรม CNN ส่วนใหญ่นั้น แต่เดิมมีขนาดใหญ่เกินไปที่จะใช้ในอุปกรณ์เคลื่อนที่ Raspberry Pi หรือ IoT ได้แบบเรียลไทม์ จึงได้คิดค้นวิธีเพื่อลดจำนวนพารามิเตอร์โมเดล Xception กลายเป็น CNN ที่สามารถใช้งานได้จริง

2.5. การวัดประสิทธิภาพ (Performance Evaluation)

การวัดประสิทธิภาพโมเดลของ (Data Mining) การที่จะนำ Model(โมเดล) ไปใช้งานจริงได้นั้น จำเป็นต้องมีการวัดประสิทธิภาพ Model(โมเดล) เสียก่อนว่า Model(โมเดล) นั้นมีประสิทธิภาพเพียงพอที่จะนำ Model(โมเดล) ดังกล่าวมาพัฒนา หรือนำ Model(โมเดล) ที่ได้ ไปใช้งานด้านต่างๆ ซึ่งการวัดประสิทธิภาพนั้นส่วนใหญ่จะวัดค่าจากใน Table(ตาราง) ข้อมูลที่มี

โดยทั่วไปแล้วจะมีตัววัดที่นิยมใช้กันในงานวิจัยและการทำงานต่างๆ อยู่ 3 ค่า และสมการคือ

1. Precision (พีสิชั่น) เป็นการวัดความแม่นยำของข้อมูล โดยพิจารณาแยกทีละคลาส

$$\frac{TP}{TP + FP}$$

2. Recall (รีคอล) เป็นการวัดความถูกต้องของ Model (โมเดล) โดยพิจารณาแยกทีละคลาส

$$\frac{TP}{TP + FN}$$

3. Accuracy (แอคคูเลซี่) เป็นการวัดความถูกต้องของ Model (โมเดล) โดยพิจารณารวมทุกคลาส

$$\frac{TP + TN}{TP + TN + FP + FN}$$

True Positive (ทรู โพสสิทีฟ) (TP) คือ จำนวนข้อมูลที่ทำนายถูกว่าเป็นคลาสซึ่งกำลังสนใจอยู่

True Negative (ทรู เน็กกาทีฟ) (TN) คือ จำนวนข้อมูลที่ทำนายถูกว่าเป็นคลาสซึ่งไม่ได้สนใจอยู่

False Positive (ฟอล โพสสิทีฟ) (FP) คือ จำนวนข้อมูลที่ทำนายผิดมาเป็นคลาสซึ่งกำลังสนใจอยู่

False Negative (ฟอล เน็กกาทีฟ) (FN) คือ จำนวนข้อมูลที่ทำนายผิดมาเป็นคลาสซึ่งไม่ได้สนใจอยู่

โดยตัววัดนี้จะมาใช้ กับตาราง Confusion Matrix (คอนฟูชั่น แมททริก)

4. Confusion Matrix (คอนฟูชั่น แมททริก) คือ ตารางแบบจัตุรัสโดยมีจำนวนแถวเท่ากับจำนวนคอลัมน์และเท่ากับจำนวนคลาส ดังภาพ มีคลาสคำตอบอยู่ 2 ค่า คือ yes และ no ฉะนั้นตารางนี้จะสร้างได้เป็นตารางขนาด 2x2 โดยข้อมูลด้านคอลัมน์คือ คลาสที่อยู่ในข้อมูลดาต้าของเรา และข้อมูลในแนวแถว คือ คลาสที่โมเดลทำนายมาได้ ตาราง Confusion Matrix (คอนฟูชั่น แมททริก) คลาสที่อยู่ในข้อมูลดาต้า actual (แอ็คทูอาว) และข้อมูลในแนวแถว คือ คลาสที่โมเดลทำนายมาได้ predicted (พีดิกเต็ด)

Predicted / Actual	yes	no
yes	TP	FP
no	FN	TN

No.	Outlook	Temperature	Humidity	Windy	Play
1	sunny	hot	normal	FALSE	?

ข้อมูลที่นำไปพยากรณ์ ตารางนี้จะเป็นนำข้อมูลต่างๆในตารางมาพยากรณ์ว่า Class Play (คลาสเพย์)

ตัวอย่างตารางข้อมูลที่นำไปพยากรณ์แล้ว

No	Actual	Predicted
1	no	no
2	no	no
3	yes	no
4	yes	yes
5	yes	no
6	no	yes
7	yes	yes
8	no	no
9	yes	no
10	yes	yes

True Positive(ทรู โพสสิทีฟ) (TP) คือ จำนวนข้อมูลที่ทำนายถูกว่าเป็นคลาส Play = yes มีจำนวน 3 ตัว (แถวที่เป็นตัวหนา คือ แถวที่ 4, 7 และ 10)

True Negative(ทรู เน็กกาทีฟ) (TN) คือ จำนวนข้อมูลที่ทำนายถูกว่าเป็นคลาส Play = no มีจำนวน 3 ตัว (แถวที่เป็นตัวเอียง คือ แถวที่ 1, 2 และ 8

False Positive (ฟอล โพสสิทีฟ)(FP) คือ จำนวนข้อมูลที่ทำนายผิดมาเป็นคลาส Play = yes มีจำนวน 1 ตัว (แถวที่ขีดเส้นใต้ คือ แถวที่ 6)

False Negative(ฟอล เน็กกาทีฟ) (FN) คือ จำนวนข้อมูลที่ทำนายผิดมาเป็นคลาส Play = no มีจำนวน 3 ตัว (แถวที่ตัวอักษรปกติ คือ แถวที่ 3, 5 และ 9)

ดังนั้นจึงสร้างตาราง Confusion Matrix(คอนฟูชั่น แมททริก)

การวัดประสิทธิภาพข้อมูลของตาราง Confusion Matrix(คอนฟูชั่น แมททริก)

Predicted / actual	yes	no
yes	3	1
no	3	3

ค่า Precision (พิสิชั่น) ของคลาส yes คือ Precision(พิสิชั่น) (Play=yes) = 3/4 = 75%

ค่า Recall (รีคอล) ของคลาส yes คือ Recall(รีคอล) (Play=yes) = 3/6 = 50%

ค่า Accuracy (แอคคูลูซี่) เป็นการวัดความถูกต้องของโมเดล โดยพิจารณารวมทุกคลาส คือ จำนวน True Positive (ทู โพสสิทีฟ) ของทุกคลาสรวมกันได้เท่ากับ 6/10 = 60%

2.6. งานวิจัยที่เกี่ยวข้อง

D. Viet Sang, N. Van Dat, and D. Phan Thuan, (2017) ได้ศึกษาเกี่ยวกับ การแสดงออกทางใบหน้าระหว่างมนุษย์ การมีปฏิสัมพันธ์แบบตัวต่อตัว ซึ่งสามารถรับรู้การแสดงนั้นของเครื่อง วิธีการเรียนรู้เครื่องจักรแบบคลาสสิก มีกระบวนการสกัดคุณลักษณะที่ซับซ้อนและผลิตผล ผลที่ไม่ดี ในเรื่องนี้จะใช้ความก้าวหน้าล่าสุดในด้าน Deep Convolutional Neural Networks (CNNs) ที่สามารถจำแนกการแสดงออกทางใบหน้าได้อย่างถูกต้อง ใช้ชุดข้อมูล FERC-2013 มีการทำ Data preprocessing ทำให้ข้อมูลเป็นบรรทัดฐานรูปแบบเดียวกันและทำให้เป็นภาพ แบบ normalizing ข้อมูลต่อพิกเซล และมีการทำ Data augmentation เป็นการเพิ่มข้อมูลเนื่องจากข้อมูลการฝึกอบรมจำนวนน้อย จึงใช้เทคนิคการเพิ่มข้อมูลเพื่อเพิ่มจำนวนของตัวอย่างการฝึกอบรมเพื่อหลีกเลี่ยงการ over fitting และปรับปรุงความถูกต้องของการรับรู้ สำหรับแต่ละรูปภาพเราจะดำเนินการแปลงต่อเนื่อง

การ Training ขั้นตอนการฝึกอบรมเราจะลดฟังก์ชันการสูญเสียโดยใช้การไล่ระดับสีแบบแบล็กแบทช์แบบมินิแบทช์ด้วยโมเมนตัมและอัลกอริธึมการถดถอยด้านหลัง ขนาดชุด คือ 256 โมเมนตัม คือ 0.9 เพื่อหลีกเลี่ยงการ over fitting และใช้เทคนิค dropout กับชั้นที่เชื่อมต่อกันอย่าง โดยมีค่า dropout เท่ากับ 0.5 ในระหว่างการฝึกอบรมเราใช้กลยุทธ์ที่ลดอัตราการเรียนรู้ 10 ครั้ง หากการสูญเสียการฝึกอบรมหยุดการปรับปรุง การทดลองแสดงให้เห็นว่าอัตราการเรียนรู้มักจะลดลงประมาณ 5 ครั้งและระยะเวลาการฝึกอบรมมักจะหมดไปหลังจากประมาณ 1400 รอบ โดยมีโครงการดังภาพที่ 2.28

ConvNet Configuration			
BKVGG8	BKVGG10	BKVGG12	BKVGG14
8 layers	10 layers	12 layers	14 layers
Input (42 × 42 × 1)			
conv3-32	conv3-32	conv3-32 **conv3-32**	conv3-32 conv3-32
maxpool			
conv3-64	conv3-64	conv3-64 **conv3-64**	conv3-64 conv3-64
maxpool			
conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128 **conv3-128**
maxpool			
conv3-256 conv3-256	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256	conv3-256 conv3-256 conv3-256 **conv3-256**
FC-256			
FC-256			
FC-7			

ภาพที่ 2.28 Deep Convolutional Neural Networks (CNNs)

ที่มา: D. Viet Sang, N. Van Dat, and D. Phan Thuan, (2017)

W. Wan, C. Yang, and Y. Li. (2016) ได้ศึกษาเกี่ยวกับเรื่อง เครือข่ายประสาทเทียมได้รับการรับรอง ว่าเป็นวิธีที่ทันสมัยที่สุดในงานต่างๆ ที่เกี่ยวข้อง มีการประยุกต์ใช้เครือข่ายประสาทเทียมกับใบหน้า โดยมีการเทรนโมเดลด้วยข้อมูล facial expression recognition (FER2013) ที่ถูกสร้างขึ้นมาจาก Kaggle facial expression challenge ระหว่างชุดข้อมูลและประสิทธิภาพของสถาปัตยกรรมเครือข่ายต่างๆ ซึ่งวิธีการฝึกอบรมจะใช้การฝึกอบรมที่แตกต่างกัน โดยใช้รูปแบบโมเดลของ AlexNet

และ VGGNet จากนั้นได้พัฒนาเพิ่มเติมโดยการเพิ่ม ชั้นของ Convolutional เป็น 8-layer-net และ 11-layer-net จากผลการทดลองพบว่า 11-layer-net ให้ผลที่ดีที่สุดมีความแม่นยำ 71.2% ใน 7 กลุ่ม โดยสถาปัตยกรรมของ 11-layer-net มีดังภาพที่ 2.29

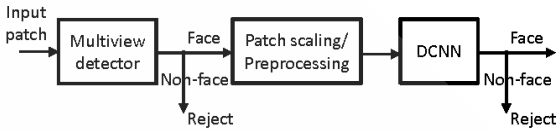
input data(48x48 grey scale image)
Data Augmentation
CONV $3 \times 3 \times 64$, RELU, BATCH NORM
CONV $3 \times 3 \times 64$, RELU, BATCH NORM
MAXPOOL 2×2
CONV $3 \times 3 \times 128$, RELU, BATCH NORM
CONV $3 \times 3 \times 128$, RELU, BATCH NORM
MAXPOOL 2×2, DROPOUT 0.2
CONV $3 \times 3 \times 256$, RELU, BATCH NORM
CONV $3 \times 3 \times 256$, RELU, BATCH NORM
MAXPOOL 2×2, DROPOUT 0.25
CONV $3 \times 3 \times 512$, RELU, BATCH NORM
CONV $3 \times 3 \times 512$, RELU, BATCH NORM
MAXPOOL 2×2, DROPOUT 0.25
FC 1024, BATCH NORM, RELU, DROPOUT 0.45
FC 1024, BATCH NORM, RELU, DROPOUT 0.45
FC 7
SOFTMAX

ภาพที่ 2.29 Architecture of our 11-layer network

ที่มา: W. Wan, C. Yang, and Y. Li. (2016)

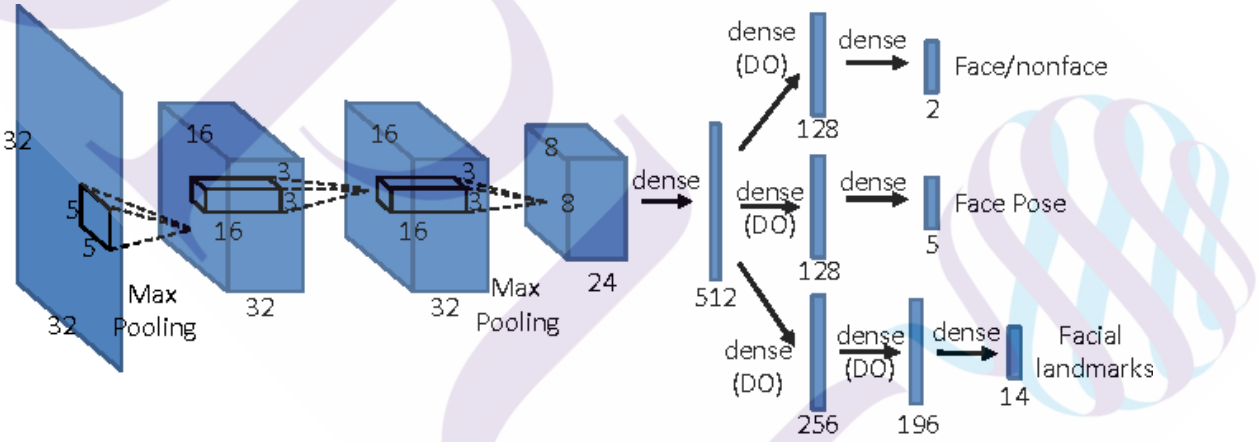
Cha Zhang and Zhengyou Zhang (2014) ได้ศึกษาเกี่ยวกับเรื่อง Improving Multiview Face Detection with Multi-Task Deep Convolutional Neural Networks พบว่า การตรวจจับใบหน้าหลายมิติเป็นปัญหาที่ท้าทายเนื่องจากการเปลี่ยนแปลงรูปร่างภายใต้สภาพท่าทางและสภาพการแสดงออกต่างๆ วิจัยเรื่องนี้นำเสนอ เทคนิควิธีการต่างๆเพื่อเพิ่มประสิทธิภาพในการตรวจจับ โดยเฉพาะอย่างยิ่งเราจะสร้างเครือข่ายประสาทเทียมแบบลึกซึ่งสามารถเรียนรู้การตัดสินใจใบหน้าและไม่ใช้ใบหน้า, การจัดรูป

ใบหน้าและการแก้ไขปัญหาการจดจำใบหน้า เราแสดงให้เห็นว่าโครงการการเรียนรู้หลายงานเช่นนี้สามารถปรับปรุงความถูกต้องของผู้ลักษณนามได้ ในชุดข้อมูล FDDB



ภาพที่ 2.30 Algorithm flow for predicting whether an image patch is a face or not

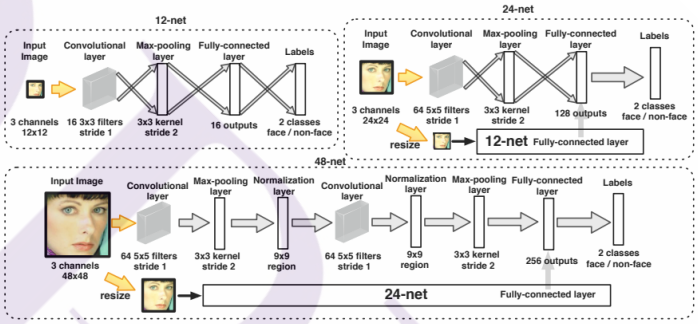
ที่มา: Cha Zhang and Zhengyou Zhang (2014)



ภาพที่ 2.31 The multi-task DCNN network adopted in this paper

ที่มา: Cha Zhang and Zhengyou Zhang (2014)

Haoxiang Li and other (2015) ได้ศึกษาเกี่ยวกับเรื่อง A Convolutional Neural Network Cascade for Face Detection เรื่องนี้เป็นการนำเสนอ Convolutional Neural Network ที่สร้างขึ้นจากการสลับซับซ้อนเครือข่ายประสาทเทียม Convolutional Neural Network (CNNs) ที่มีประสิทธิภาพมากในการจำแนกลักษณะที่มีประสิทธิภาพสูง cascade ได้ เป็นรูปแบบที่ได้รับความนิยมและมีประสิทธิภาพมากที่สุด ในการรตรวจจับใบหน้า CNN สามารถเรียนรู้คุณลักษณะต่างๆในการจับภาพที่ซับซ้อนได้ โดยอัตโนมัติ CNN สามารถเรียนรู้คุณลักษณะต่างๆ เพื่อจับภาพรูป (ภาพใบหน้า) แบบที่ซับซ้อนได้ โดยการ training data และ testing ที่มีจำนวนมาก ในการประมวลผลจะใช้ GPU เพื่อเพิ่มความเร็วในการประมวลผล

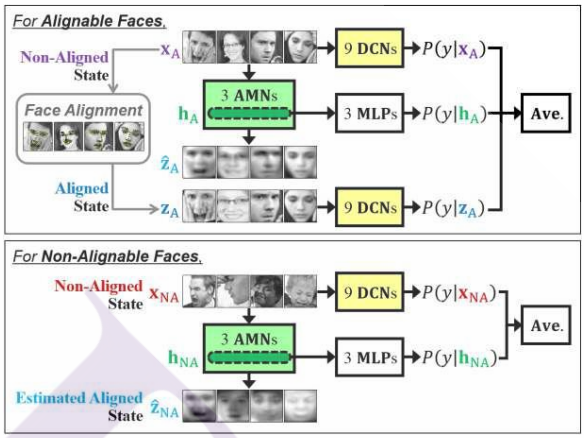


ภาพที่ 2.32 CNN structures of the 12-net, 24-net and 48-net

ที่มา: Haoxiang Li and other (2015)

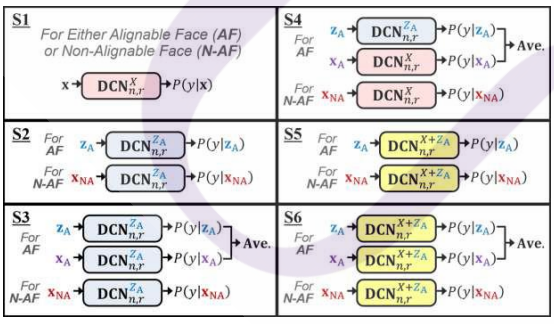
Bo-Kyeong Kim, Suh-Yeon Dong, Jihyeon Roh, Geonmin Kim and Soo-Young Lee (2016) ได้ศึกษาเกี่ยวกับเรื่อง การจัดตำแหน่งในสถานการณ์จริงหรือใบหน้าจริงๆ อาจล้มเหลวได้ ซึ่งทำให้ส่งผลเสียต่อประสิทธิภาพของระบบการรับรู้การแสดงออกทางใบหน้า (FER) อัตโนมัตินี้ ในการศึกษาครั้งนี้เราสมมติสถานการณ์จริงรวมถึงใบหน้าที่ไม่สามารถจัดตำแหน่งได้เนื่องจากความล้มเหลวในการตรวจหาจุดสังเกตใบหน้า วิธีการที่เราเสนอไว้นี้จะทำให้ข้อมูลเกี่ยวกับสถานะใบหน้าที่ไม่สอดคล้องและสอดคล้องกันเพื่อเพิ่มความแม่นยำและประสิทธิภาพของระบบการรับรู้การแสดงออกทางใบหน้า FER มีการจำลองสถานการณ์ทั้งหมด 6 สถานการณ์ โดยใช้ deep Convolutional neural networks (DCNs) และวิเคราะห์สาเหตุของความแตกต่างของประสิทธิภาพ เพื่อให้ใบหน้าที่สามารถจัด

ตำแหน่งได้ดีขึ้น นอกจากนี้เรายังแนะนำ DCN ที่เรียนรู้การทำ alignment-mapping networks (AMNs) ใบหน้าที่ไม่ได้จัดชิดให้เป็นแนวเดียวกัน แสดงให้เห็นว่า DCN และ AMN ทำให้ได้ผลลัพธ์ที่ดีสำหรับ FER



ภาพที่ 2.33 Our automatic FER system contains several DCNs

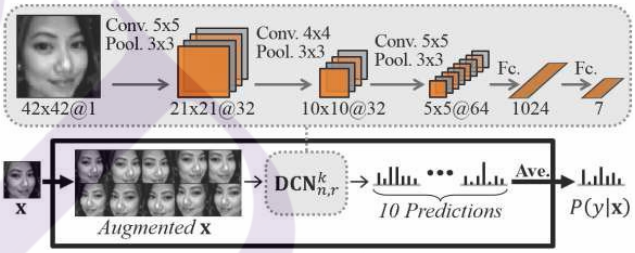
ที่มา: Bo-Kyeong Kim, Suh-Yeon Dong, Jihyeon Roh, Geonmin Kim and Soo-Young Lee (2016).



ภาพที่ 2.34 สถานการณ์ 6 รูปแบบในการรวมข้อมูลแบบ aligned และ non-aligned

ที่มา: Bo-Kyeong Kim, Suh-Yeon Dong, Jihyeon Roh, Geonmin Kim and Soo-Young Lee (2016).

โดยในการนเสนอแนวทางนี้มีการทำ Ensemble ในการใช้ deep Convolutional neural networks (DCNs) กับข้อมูล FER2013 นี้ ซึ่งในช่วงการเรียนรู้เลือกปฏิบัติ DCNs จะเป็นประโยชน์ในการใช้ชุดข้อมูลการฝึกอบรมที่ผสานของใบหน้าที่ alignable และ non-alignable ได้เพื่อประสิทธิภาพในการประเมินผลตลอดจนประสิทธิภาพในการปฏิบัติงานของ FER จากนั้นทำการวิเคราะห์ผลจากการทดลอง ในขั้นตอนการทดสอบสำหรับการปรับแนวผิวหน้าการรวม ข้อมูลของ non-alignable และ DCNs ที่เลือกปฏิบัติช่วยเพิ่มความถูกต้องของ FERในขั้นตอนการทดสอบสำหรับใบหน้าที่ไม่สามารถจัดตำแหน่งและปรับตำแหน่งได้ควรเพิ่มข้อมูลในระดับการตัดสินใจจากการจัดองค์ประกอบที่ซ่อนอยู่ของ AMN

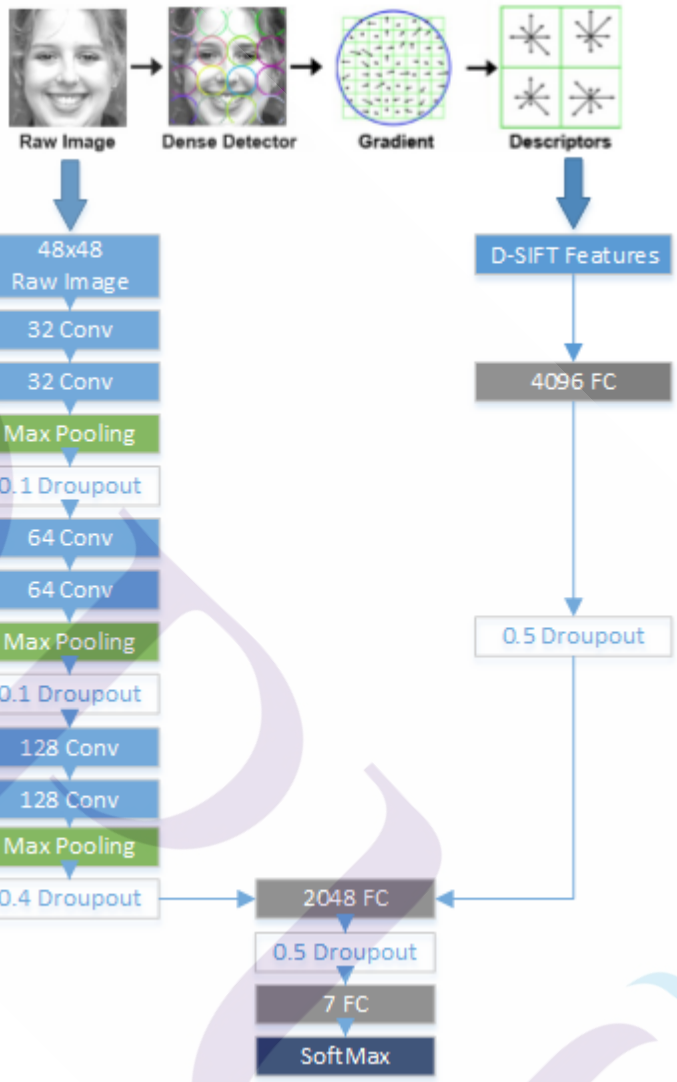


ภาพที่ 2.35 สถาปัตยกรรมของ DCN และขั้นตอนการประเมินผลด้วยการเพิ่มข้อมูล

ที่มา: Bo-Kyeong Kim, Suh-Yeon Dong, Jihyeon Roh, Geonmin Kim and Soo-Young Lee (2016).

ทำให้มีความหลากหลายในการเลือกใช้โมเดลในการเทรนข้อมูลในแต่ละชุดที่เหมือนกันเพื่อนำมาเทรนและสร้างโมเดล Classify ที่หลากหลายรูปแบบ และสุดท้ายนำผมลัพธ์ของแต่ละโมเดลมาเปรียบเทียบแล้วเลือกค่าผลลัพธ์ของโมเดลที่ดีที่สุดออกมาใช้ โดยรูปแบบสถานการณ์ที่ดีที่สุดคือรูปแบบ S6 ซึ่งโมเดลนี้ใช้ชุดข้อมูลการฝึกอบรมที่ผสานของ $X + Z_A$ จะเป็นประโยชน์ในสิ่งที่สามารถเรียนรู้เรื่องความรู้ที่หลอมรวมกันเกี่ยวกับสถานะใบหน้าที่ไม่เรียงชิดและชิดกันได้ ได้รับการยืนยันโดยการปรับปรุงความถูกต้องทั้ง AF และ N-AF ใน S6 เมื่อผ่านกระบวนการต่างๆ ทำให้ได้ค่า Ensemble เท่ากับ 73.31%

M. Al-Shabi, W. Ping Cheah and T. Connie (2017) ได้ศึกษาเกี่ยวกับ องค์ประกอบการจดจำใบหน้าที่มีประสิทธิภาพ เป็นสิ่งสำคัญสำหรับระบบปฏิสัมพันธ์ของมนุษย์และคอมพิวเตอร์ ซึ่งแนวทางใหม่ในการจดจำใบหน้า จะเน้นที่ความถูกต้องแม่นยำ ในขณะที่ข้อมูลตัวอย่างที่ใช้ในการวัดผลมีเพียงเล็กน้อยสำหรับการฝึกอบรม จึงต้องใช้การศึกษาคุณลักษณะ (SIFT) ของสเกล ถูกนำมาใช้เพื่อเพิ่มประสิทธิภาพให้กับข้อมูลขนาดเล็ก เนื่องจาก SIFT ไม่ต้องการข้อมูลการฝึกอบรมที่มีขนาดใหญ่หรือจำนวนมาก ในที่นี้จะศึกษาและเปรียบเทียบ SIFT แบบปกติเมื่อผสานเข้ากับคุณลักษณะ CNN นอกจากนี้ แนวทางที่เสนอจะได้รับการทดสอบในชุดข้อมูล FER-2013 และ CK + ซึ่งประสบความสำเร็จ 73.4% ใน FER-2013 และ 99.1% ใน CK + ดังตัวอย่างโมเดล ภาพที่ 2.36

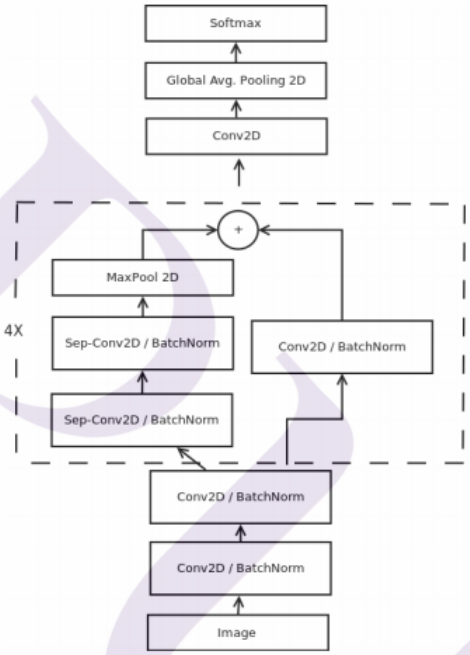


ภาพที่ 2.36 CNN-SIFT Hybrid method

ที่มา: M. Al-Shabi, W. Ping Cheah and T. Connie (2017)

Octavio Arriaga and Paul G. Ploger (2017) ได้ศึกษาเกี่ยวกับเรื่อง Real-time Convolutional Neural Networks for Emotion and Gender Classification ได้สร้างการตรวจหาใบหน้าแบบเรียลไทม์

พร้อมกับการจำแนกเพศและการจำแนกอารมณ์ โดยใช้เครือข่ายประสาทศัลยศาสตร์ Convolutional Neural Network (CNN) เพื่อใช้งานในหุ่นยนต์ Care-O-bot 3 โดยมีการแบ่งอารมณ์ความรู้ออกเป็น 7 รูปแบบ ได้แก่ angry, disgust, fear, happy, sad, surprise, และ neutral โดยใช้ dataset ชื่อว่า FER2013 และแยกเพศ เป็น เพศชายและเพศหญิงโดยใช้ dataset ชื่อว่า IMDB ดำเนินการโดยการทำ convolution 9 layers ใช้ ReLU activations และ Average Pooling มีจำนวน parameters เท่ากับ 600,000 ค่าความถูกต้องของ FER-2013 เท่ากับ 66% และ IMDB มีค่าความถูกต้อง เท่ากับ 96%

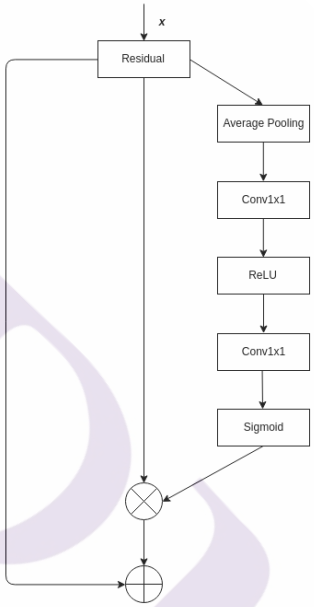


ภาพที่ 2.37 Our proposed model for real-time classification.

ที่มา: Octavio Arriaga and Paul G. Ploger (2017)

Andrinandrasana David Rasamoelina, Fouzia Adjailia and Peter SINC AK´. (2019) ได้ศึกษาเกี่ยวกับเรื่อง Deep Convolutional Neural Network For Robust Facial Emotion Recognition อารมณ์มีบทบาทสำคัญในการสื่อสารสังคม การติดต่อและการตัดสินใจ อารมณ์ก็ยังมีความจำเป็นหรือ

เป็นหัวใจสำคัญในการสื่อสาร โดยในงานวิจัยนี้ ใช้เทคนิคการเรียนรู้ที่ล้ำลึกเพื่อจำแนกใบหน้าอย่างมีประสิทธิภาพการแสดงออก เราทดสอบอัลกอริทึมของเรากับชุดข้อมูลหลัก 4 ชุด: FER2013, AffectNet, RaFD และ KDEF และเปรียบเทียบได้รับความแม่นยำก จากการทดสอบทั้งหมดที่ดำเนินการวิธีที่เสนอนั้นทำได้ดีกว่าโซลูชั่นที่มีอยู่ การศึกษาครั้งนี้มุ่งเน้นการแสดงออกทางสีหน้า ซึ่งโครงสร้างโมเดลมีลักษณะดังภาพ

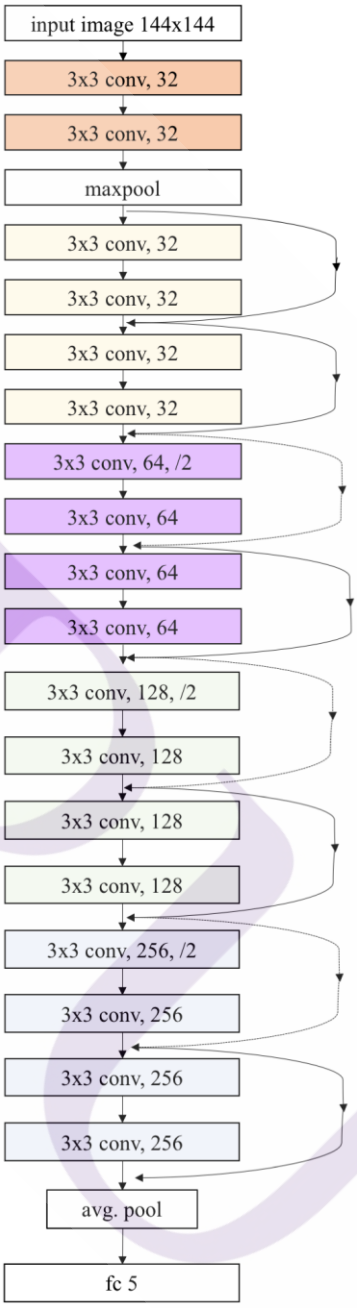


ภาพที่ 2.38 Residual Squeeze and Excitation block.

ที่มา: Andrinandrasana David Rasamoelina, Fouzia Adjailia and Peter SINC ̆ AK ́. (2019)

จากภาพที่ 2.37 สถาปัตยกรรมนี้มีการผสมผสานการใช้งานของโมดูลต่างๆ คือ Residual, squeeze และ excitation โดย Residual โดย Residual เป็นการปรับเปลี่ยนการทำแผนที่ที่ต้องการระหว่างสองเลเยอร์ที่ตามมาเพื่อให้ฟีเจอร์ที่เรียนรู้กลายเป็นความแตกต่างของแผนที่คุณลักษณะดั้งเดิมและคุณสมบัติที่ต้องการ โมดูล Squeeze and Excitation เป็บล็อกสำหรับ CNN ที่ปรับปรุงการพึ่งพาซึ่งกันและกัน วัตถุประสงค์หลักของการนี้ คือ การเพิ่มพารามิเตอร์ในแต่ละช่องทางของ convolutional เพื่อให้เครือข่ายสามารถปรับน้ำหนักได้อย่างเหมาะสม

Natalia Efremova, Mikhail Patkin and Denis Sokolov. (2019) ได้ศึกษาเกี่ยวกับเรื่อง Face and Emotion Recognition with Neural Networks on Mobile Devices: Practical Implementation on Different Platforms โดยเปรียบเทียบประสิทธิภาพของแบบจำลองในแพลตฟอร์มต่างๆ ได้แก่ GPU, mobile และ Raspberry Pi โดยใช้ pre-trained จากโมเดล FaceNet ซึ่งในการออกแบบโครงสร้าง CNN มีการใช้ residual มีจำนวน 20 parametrized ดังภาพ



ภาพที่ 2.39 residual network with 20 parametrized layers

ที่มา: Natalia Efremova, Mikhail Patkin and Denis Sokolov. (2019)

มีการทำ data augmentation ชุดข้อมูลถูกรวมเข้าด้วยกันจากสองแหล่ง (1) แหล่งข้อมูลสาธารณะเช่นข้อมูลการค้นหาของ Google (การค้นหารูปภาพคงที่, gif และวิดีโอจากภาพยนตร์) และวิดีโอ YouTube และ(2) การบันทึกวิดีโอจากผู้เข้าร่วมระหว่างการสาธิตวิดีโอคลิป คนสี่สิบแปดคนเข้าร่วมในการศึกษานี้ (หญิง 20 คนอายุระหว่าง 19-68 ปี)

บทความนี้นำเสนออารมณ์ข้ามแพลตฟอร์มที่แปลกใหม่ สถาปัตยกรรมการรับรู้แบบจำลองที่เสนอนั้นสามารถที่จะเรียนรู้อารมณ์ จำนวน 5 กลุ่มอารมณ์ ผลการดำเนินงานของเราแอปพลิเคชันการจดจำอารมณ์สามารถเปรียบเทียบกับประสิทธิภาพของเครือข่ายประสาทเทียมที่ทันสมัย ระบบต้องการอุปกรณ์พกพาที่มีความสำคัญความจุในการคำนวณ (iPhone 7 ขึ้นไป) และอุปกรณ์ที่ฝังตัว Raspberry Pi (Movidius)

จากการเปรียบเทียบประสิทธิภาพของแบบจำลองที่แตกต่างกันแพลตฟอร์มระหว่างงานการจดจำใบหน้าและอารมณ์ ความแม่นยำของการจดจำใบหน้าบนแพลตฟอร์มมือถือคือ 98%

บทที่ 3
ระเบียบวิธีวิจัย

วิจัยเรื่องนี้มีวัตถุประสงค์เพื่อศึกษาการจำแนกอารมณ์จากใบหน้าโดยใช้ โครงสร้างเครือข่ายประสาทเทียม (Convolutional Neural Networks: CNNs) และเพื่อจัดกลุ่มการแสดงออกทางสีหน้า 7 กลุ่ม ซึ่งเป็นวิจัยเชิงทดลอง (Experimental Research) โดยจะเป็นการนำเสนอโมเดลที่มีการแยกชั้นที่ลึกโดยอ้างอิงรูปแบบมาจาก Xception Model ดังนั้นสมมุติฐานของแบบจำลองนี้ คือ การทำแผนที่ความสัมพันธ์ระหว่างช่องทางและความสัมพันธ์เชิงพื้นที่ โดยมีขั้นตอนการออกแบบให้สถาปัตยกรรมเครือข่ายนี้ มีความซับซ้อนที่น้อยลงกว่า Xception Model จากาการใช้ชุดข้อมูล FER2013 เพื่อเพิ่มและรักษาความแม่นยำในการจำแนก

3.1 ขั้นตอนวิธีการดำเนินการวิจัย

วิจัยเรื่องนี้มีวัตถุประสงค์เพื่อศึกษาการจำแนกอารมณ์จากใบหน้าโดยใช้ โครงสร้างเครือข่ายประสาทเทียม (Convolutional Neural Networks: CNNs) และเพื่อจัดกลุ่มการแสดงออกทางสีหน้า 7 กลุ่ม ซึ่งเป็นวิจัยเชิงทดลอง (Experimental Research) โดยจะเป็นการนำเสนอโมเดลที่มีการแยกชั้นที่ลึกโดยอ้างอิงรูปแบบมาจาก Xception Model ดังนั้นสมมุติฐานของแบบจำลองนี้ คือ การทำแผนที่ความสัมพันธ์ระหว่างช่องทางและความสัมพันธ์เชิงพื้นที่ โดยมีขั้นตอนการออกแบบให้สถาปัตยกรรมเครือข่ายนี้ มีความซับซ้อนที่น้อยลงกว่า Xception Model จากาการใช้ชุดข้อมูล FER2013 เพื่อเพิ่มและรักษาความแม่นยำในการจำแนก

3.1.1 ศึกษาข้อมูลเกี่ยวกับ FACS (Facial Action Coding System)

3.1.2 ออกแบบโครงสร้างเครือข่ายประสาทเทียม (Convolutional Neural Networks: CNNs) ปรับปรุงแบบมาจาก Xception Model

3.1.3 การสร้างโมเดลสำหรับการจำแนกอารมณ์

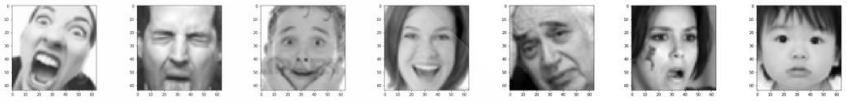
3.1.3.1 โมเดลสำหรับการจำแนกอารมณ์ 7 กลุ่มอารมณ์

3.1.3.2 โมเดลสำหรับการจำแนกอารมณ์ 2 กลุ่มอารมณ์

3.1.4 วิเคราะห์และสรุปผล

3.1.1 ศึกษาข้อมูลเกี่ยวกับ FACS (Facial Action Coding System)

ศึกษาข้อมูลที่เกี่ยวกับการแสดงอารมณ์ออกทางสีหน้าในระหว่างการปฏิสัมพันธ์ระหว่างมนุษย์ จากงานวิจัยเรื่องนี้จะมุ่งเน้น ที่ใช้การเรียนรู้ชุดข้อมูล FER-2013 ที่ถูกสร้างขึ้นโดย Kaggle ในปี 2013 ซึ่งมีการจำแนกจำนวนกลุ่มไว้ทั้งหมด 7 กลุ่ม ได้แก่ 0: angry, 1: disgust, 2: fear, 3: happy, 4: sad, 5: surprise, และ 6: neutral ดังภาพที่ 3.1



ภาพที่ 3.1 ตัวอย่างข้อมูลในข้อมูลชุด FER-2013

ตารางที่ 3.1 การแบ่งจำนวนข้อมูล FER-2013 สำหรับการสร้างโมเดลเครือข่ายประสาทเทียม (Convolutional Neural Networks: CNNs) จำแนก 7 กลุ่มอารมณ์

กลุ่มที่บ่งบอกลักษณะ	รายละเอียดข้อมูล	จำนวนใบหน้า	ข้อมูลสำหรับเทรนโมเดล	ข้อมูลสำหรับทดสอบ
Unpleasantness	ใบหน้าโกรธ	4,953	4,462	491
Unpleasantness	ใบหน้าไม่ชอบ	547	492	55
Unpleasantness	ใบหน้ากลัว (fear)	5,121	4,593	528
Unpleasantness	ใบหน้าเศร้า (sad)	6,077	5,483	594
Pleasantness	ใบหน้ามีความสุข	8,989	8,110	879
Pleasantness	ใบหน้าประหลาดใจ	4,002	3,586	416
Pleasantness	ใบหน้าแบบปกติ	6,198	5,572	626
รวมทั้งหมด		**35,887**	**32,298**	**3,589**

จากตารางที่ 3.1 การแบ่งจำนวนข้อมูล FER-2013 สำหรับการสร้างโมเดลเครือข่ายประสาทเทียม (Convolutional Neural Networks: CNNs) พบว่า ในการแบ่งจำนวนข้อมูล FER-2013 สำหรับการสร้างโมเดลเครือข่ายประสาทเทียม (Convolutional Neural Networks: CNNs) อ้างอิงรูปแบบมาจาก Xception Model แบ่งในอัตราส่วน 90:10 จากจำนวนข้อมูลทั้งหมด 35,887

ภาพ โดยชุดที่ใช้สำหรับสร้างโมเดลมีจำนวน 32,298 ภาพ และชุดที่ใช้สำหรับการทดสอบมีจำนวน 3,589 ภาพ ตามลำดับ

ข้อมูลของโมเดลจำแนก Pleasantness/Unpleasantness จะประกอบได้ด้วย 2 กลุ่ม ได้แก่ กลุ่มที่บ่งบอกลักษณะว่ามีความสนใจ (Pleasantness) กลุ่มที่บ่งบอกลักษณะว่าไม่มีความสนใจ (Unpleasantness)

กลุ่มที่บ่งบอกลักษณะว่ามีความสนใจ (Pleasantness) ได้แก่ ใบหน้ามีความสุข, ใบหน้าประหลาดใจ, และ ใบหน้าแบบปกติ

กลุ่มที่บ่งบอกลักษณะว่าไม่มีความสนใจ (Unpleasantness) ได้แก่ ใบหน้าโกรธ, ใบหน้าไม่ชอบ, ใบหน้าเศร้า, และ ใบหน้ากลัว

ตารางที่ 3.2 การแบ่งจำนวนข้อมูล FER-2013 สำหรับการสร้างโมเดลเครือข่ายประสาทเทียม (Convolutional Neural Networks: CNNs) จำแนก 2 กลุ่มอารมณ์ (Pleasantness/Unpleasantness)

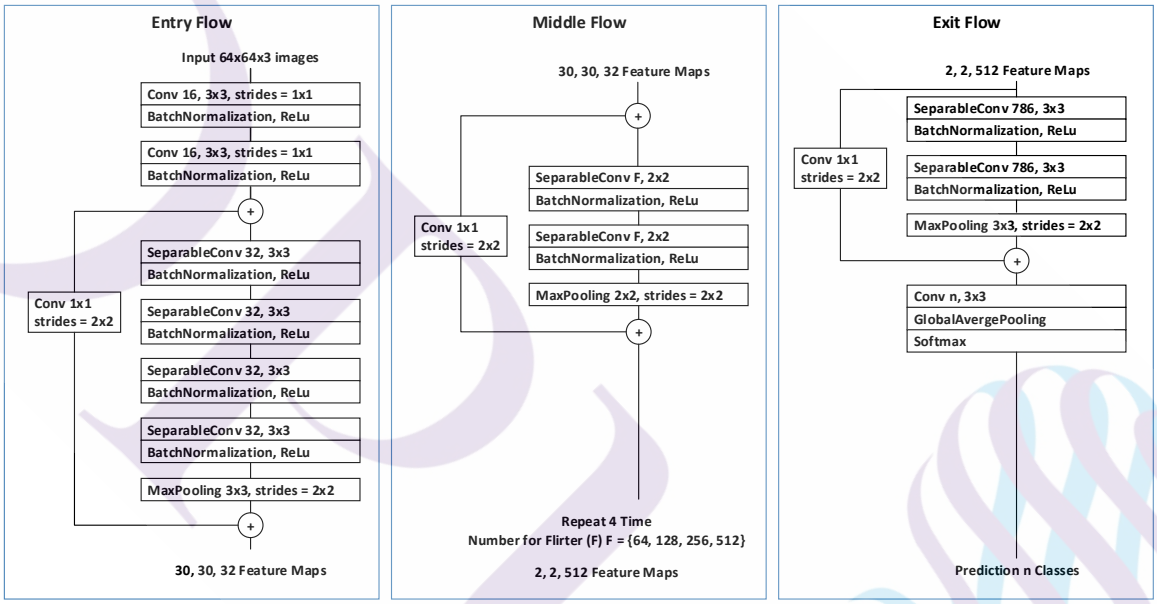
รายละเอียดข้อมูล	**จำนวนใบหน้า**	**ข้อมูลสำหรับเทรนโมเดล**	**ข้อมูลสำหรับทดสอบโมเดล**
กลุ่มที่บ่งบอกลักษณะว่ามีความสนใจ (Pleasantness)	19,189	17,268	1,921
กลุ่มที่บ่งบอกลักษณะว่าไม่มีความสนใจ (Unpleasantness)	16,698	15,030	1,668
รวมทั้งหมด	**35,887**	**19,189**	**3,589**

จากตารางที่ 3.2 ข้อมูลกลุ่มที่บ่งบอกลักษณะว่ามีความสนใจ (Pleasantness) และกลุ่มที่บ่งบอกลักษณะว่าไม่มีความสนใจ (Unpleasantness) พบว่า จำนวนใบหน้าทั้งหมดมีจำนวน 35,887 ภาพ โดยแบ่งเป็นกลุ่มที่บ่งบอกลักษณะว่ามีความสนใจ (Pleasantness) มีจำนวนทั้งหมด 19,189 ภาพ โดยแบ่งเป็นข้อมูลสำหรับเทรนโมเดล จำนวน 17,268 ภาพ ข้อมูลสำหรับทดสอบโมเดล จำนวน 1,921 ภาพ และกลุ่มที่บ่งบอกลักษณะว่าไม่มีความสนใจ (Unpleasantness) จำนวนทั้งหมด

16,698 ภาพ โดยแบ่งเป็นข้อมูลสำหรับเทรนโมเดล จำนวน 15,030 ภาพ ข้อมูลสำหรับทดสอบโมเดล จำนวน 1,668 ภาพ ตามลำดับ

3.1.2 ออกแบบโครงสร้างเครือข่ายประสาทเทียม (Convolutional Neural Networks: CNNs) ปรับปรุงรูปแบบมาจาก Xception Model

ในงานวิจัยนี้จะนำเสนอโครงสร้างเครือข่ายประสาทเทียม (convolutional neural networks: CNNs) โดยจะเป็นการนำเสนอโมเดลที่มีการแยกชั้นที่ลึกโดยอ้างอิงรูปแบบมาจากโมเดล Xception ดังนั้นสมมุติฐานของแบบจำลองนี้ คือ การทำแผนที่ความสัมพันธ์ระหว่างช่องทางและความสัมพันธ์เชิงพื้นที่ โดยออกแบบอ้างอิงจาก Xception model



ภาพที่ 3.2 สถาปัตยกรรมแบบจำลองปรับปรุงจาก Xception model

จากภาพสถาปัตยกรรมนี้จะประกอบไปด้วย 3 ส่วน ได้แก่ ส่วนที่ 1 Entry Flow ส่วนที่ 2 Middle Flow และส่วนที่ 3 Exit Flow โดยเริ่มจากการนำข้อมูลเข้า การทำงานแบบวนซ้ำ และการจำแนกการแสดงออกทางสีหน้า ทั้งหมด n กลุ่ม

ในส่วนที่ 1 Entry Flow จะเป็นการกรองคุณสมบัติแบบหยาบๆ โดยดึงออกมาจากชั้น convolutions จำนวน 4 ชั้น โดยการทำ normalization และการใช้งานฟังชั่น activation relu ในการทำแบบนี้จะทำให้จำนวนตัวกรอง (filters) เพิ่มจำนวนขึ้นเป็นสองเท่าจาก convolutions เดิม

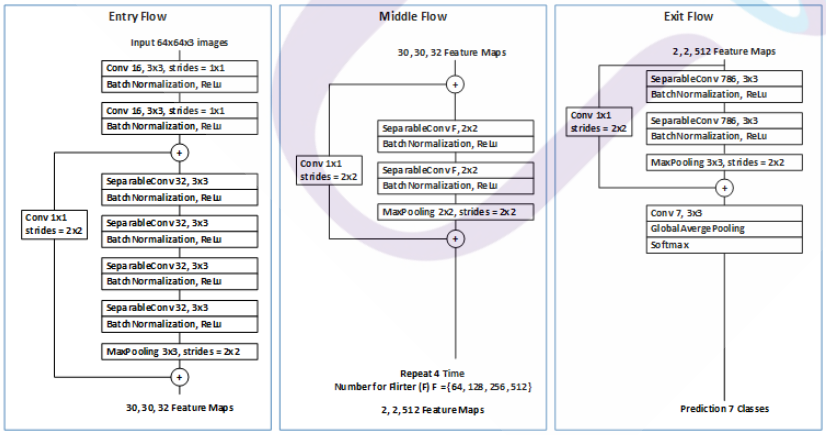
ในส่วนที่ 2 Middle Flow จะได้คุณลักษณะที่ละเอียดและซับซ้อนมากขึ้นโดยจะถูกแยก convolutions จำนวน 8 ชั้น โดยจะต้องผ่าน การแยกของ convolutions จำนวน 2 ชั้น จำนวนตัวกรอง (filters) จะเพิ่มขึ้นทละสองเท่าโดยเริ่มจาก 64, 128, 256, และ 512 เป็นตัวสุดท้ายตามลำดับ และกำหนดขนาดในการกรองเท่ากับ 2x2 mapping

ในส่วนที่ 3 Exit Flow เป็นขั้นตอนที่แสดงคุณสมบัติหรือคุณลักษณะของภาพที่มีรายละเอียดมากที่สุด ซึ่งจะถูกแยกออกเป็น 2 ชั้น โดยมีตัวกรอง (filters) คือ 786 ตัว และชั้น convolutions ชั้นสุดท้ายเท่ากับ n ชั้น มีตัวกรอง (filters) เท่ากับ 3x3 ซึ่งจะเกี่ยวข้องกับการจำแนกการแสดงอกอทางสีหน้าจำนวน n กลุ่ม และใช้ global average pooling เพื่อลดขนาดการ mapping เท่ากับ 3x3 ไปจนถึง 1x1 สุดท้ายจะใช้ฟังก์ชั่น softmax เพื่อกระตุ้นการทำงาน ในการจำแนกการแสดงออกทางสีหน้าให้ได้ถูกต้องมากที่สุด

3.1.3 การสร้างโมเดลสำหรับการจำแนกอารมณ์

3.1.3.1 โมเดลสำหรับการจำแนกอารมณ์ 7 กลุ่มอารมณ์

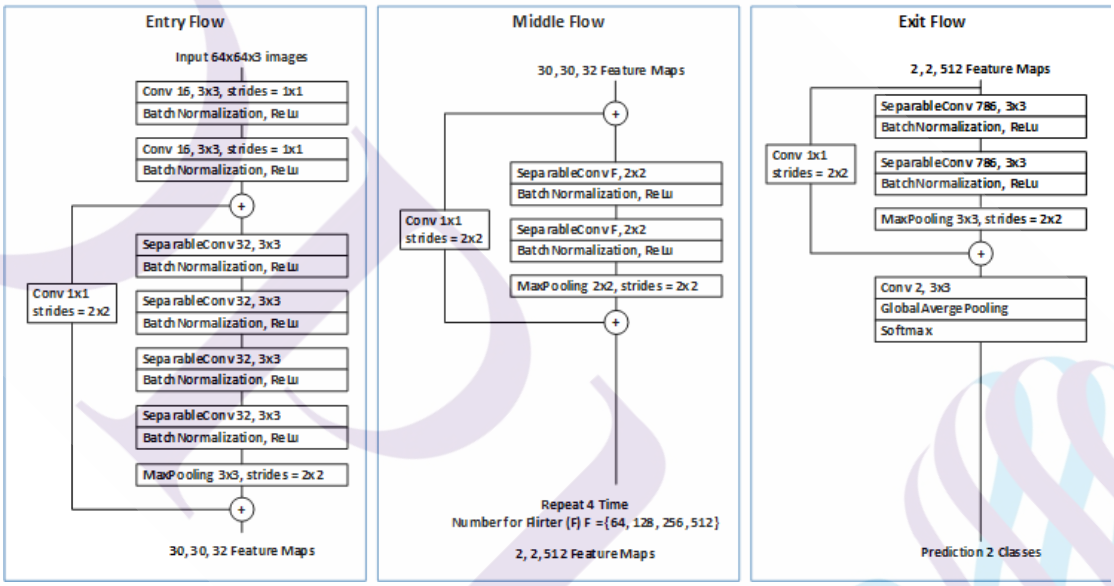
ในการปรับปรุงโมเดลสำหรับการจำแนกอารมณ์ 7 กลุ่มอารมณ์ มีการแก้ไข ในส่วนที่ 3 Exit Flow เป็นขั้นตอนที่แสดงคุณสมบัติหรือคุณลักษณะของภาพที่มีรายละเอียดมากที่สุด ซึ่งจะถูกแยกออกเป็น 2 ชั้น โดยมีตัวกรอง (filters) คือ 786 ตัว และชั้น convolutions ชั้นสุดท้ายเท่ากับ 7 ชั้น มีตัวกรอง (filters) เท่ากับ 3x3 ซึ่งจะเกี่ยวข้องกับการจำแนกการแสดงอกอทางสีหน้าจำนวน n กลุ่ม และใช้ global average pooling เพื่อลดขนาดการ mapping เท่ากับ 3x3 ไปจนถึง 1x1 สุดท้ายจะใช้ฟังก์ชั่น softmax เพื่อกระตุ้นการทำงาน ในการจำแนกการแสดงออกทางสีหน้าให้ได้ถูกต้องมากที่สุด



ภาพที่ 3.3 โมเดลสำหรับการจำแนกอารมณ์ 7 กลุ่มอารมณ์

3.1.3.2 โมเดลสำหรับการจำแนกอารมณ์ 2 กลุ่มอารมณ์

ในการปรับปรุงโมเดลสำหรับการจำแนกอารมณ์ 2 กลุ่มอารมณ์ มีการแก้ไข ในส่วนที่ 3 Exit Flow เป็นขั้นตอนที่แสดงคุณสมบัติหรือคุณลักษณะของภาพที่มีรายละเอียดมากที่สุด ซึ่งจะถูกแยกออกเป็น 2 ชั้น โดยมีตัวกรอง (filters) คือ 786 ตัว และชั้น convolutions ชั้นสุดท้ายเท่ากับ 2 ชั้น มีตัวกรอง (filters) เท่ากับ 3x3 ซึ่งจะเกี่ยวข้องกับการจำแนกการแสดงอกอทางสีหน้าจำนวน n กลุ่ม และใช้ global average pooling เพื่อลดขนาดการ mapping เท่ากับ 3x3 ไปจนถึง 1x1 สุดท้ายจะใช้ฟังก์ชั่น softmax เพื่อกระตุ้นการทำงาน ในการจำแนกการแสดงออกทางสีหน้าให้ได้ถูกต้องมากที่สุด



ภาพที่ 3.4 โมเดลสำหรับการจำแนกอารมณ์ 2 กลุ่มอารมณ์

3.1.4 วิเคราะห์และสรุปผล

เปรียบเทียบผลการทดสอบโดยใช้ค่า Accuracy, Precision, Recall และ Confusion Matrix

บทที่ 4
ผลการศึกษา

ผลการวิเคราะห์จากขั้นตอนการดำเนินการวิจัย มีดังนี้

4.1 การวัดความแม่นยำของโมเดลบนเครื่องคอมพิวเตอร์สมรรถนะสูง

4.1.1 ความแม่นยำของโมเดล จำแนกอารมณ์ 7 กลุ่มอารมณ์

4.1.2 ความแม่นยำของโมเดล จำแนกอารมณ์ 2 กลุ่มอารมณ์

4.2 เวลา/ทรัพยากรที่ใช้ในการจำแนกอารมณ์บนอุปกรณ์ฝังตัว (Raspberry Pi)

4.2.1 เวลา/ทรัพยากรที่ใช้ของโมเดล จำแนกอารมณ์ 7 กลุ่มอารมณ์

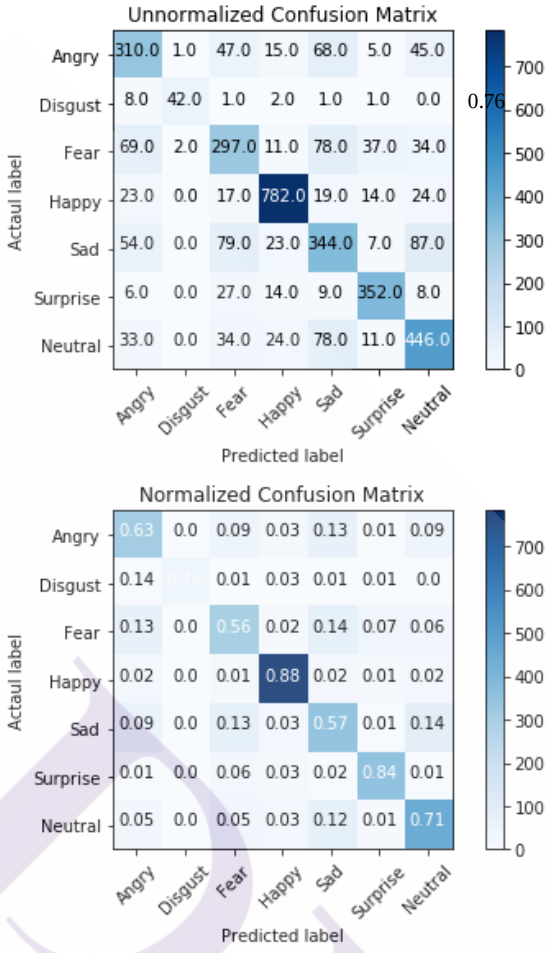
4.2.2 เวลา/ทรัพยากรที่ใช้ของโมเดล จำแนกอารมณ์ 2 กลุ่มอารมณ์

4.1 การวัดความแม่นยำของโมเดลบนเครื่องคอมพิวเตอร์สมรรถนะสูง

4.1.1 ความแม่นยำของโมเดล จำแนกอารมณ์ 7 กลุ่มอารมณ์

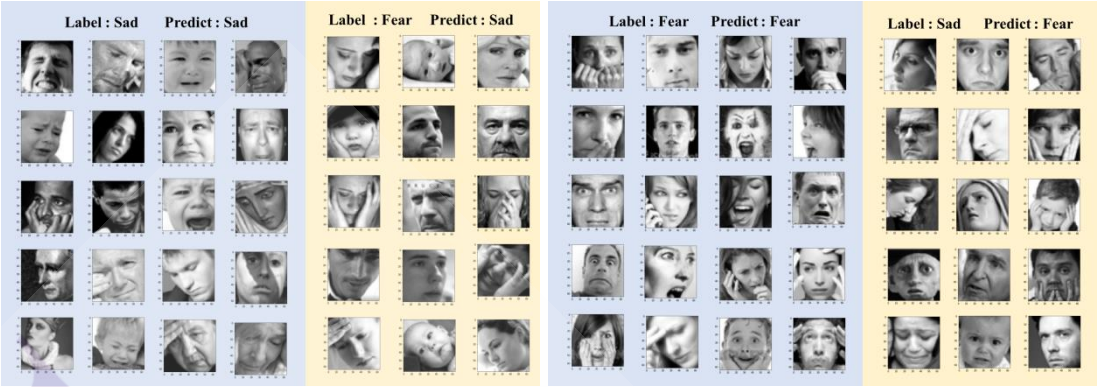
มีการจำแนกจำนวนกลุ่มไว้ทั้งหมด 7 กลุ่ม ได้แก่ ใบหน้าโกรธ, ใบหน้าไม่ชอบ, ใบหน้ากลัว, ใบหน้ามีความสุข, ใบหน้าเศร้า, ใบหน้าประหลาดใจ และใบหน้าแบบปกติ

พบว่า มีความถูกต้องในภาพรวม (Accuracy) คิดเป็นร้อยละ 71.69 สามารถจำแนกลักษณะการแสดงอารมณ์จากใบหน้ามีความสุข มีความถูกต้องมากที่สุด ซึ่งมีค่า ดังภาพที่ 4.1



ภาพที่ 4.1 ผลการออกแบบโครงสร้างเครือข่ายประสาทเทียม (Convolutional Neural Networks: CNNs)

เมื่อพิจารณากลุ่มที่มีความถูกต้องน้อยที่สุดพบว่าเป็นกลุ่มของใบหน้ากลัว (fear) และใบหน้าเศร้า ที่มีความถูกต้องเพียง 57-60% เท่านั้น จากการสังเกตภาพต้นฉบับที่ผ่านการจำแนกทำให้สามารถอธิบายได้ว่ากลุ่มของใบหน้ากลัว และใบหน้าเศร้า มีลักษณะที่ใกล้เคียงและคล้ายคลึงกันมาก ซึ่งเป็นการยากที่จะจำแนกหรือแบ่งแยกทั้ง 2 กลุ่มนี้ได้ชัดเจน ดังภาพที่ 4.2



ภาพที่ 4.2 ภาพตัวอย่างกลุ่มของใบหน้ากลัว (fear) และใบหน้าเศร้า (sad)

ตารางที่ 4.1 ผลการเปรียบเทียบโมเดลจำแนกอารมณ์ 7 กลุ่มอารมณ์

โมเดล จำแนกอารมณ์ 7 กลุ่มอารมณ์	alpha	acc	precision	recall	fbeta_score	fmeasure	Total params/M	Total FLOPS/M
simple_CNN	-	63.33	75.83	50.63	48.82	60.49	0.64	254.90
simpler_CNN	-	63.83	76.86	51.49	61.47	61.47	0.56	76.76
tiny_XCEPTION	-	63.36	74.77	51.35	60.67	60.67	0.02	6.81
mini_XCEPTION	-	66.20	75.29	56.95	64.67	64.67	0.06	20.33
big_XCEPTION	-	66.87	74.65	58.93	65.73	65.73	0.21	131.54
tiny_Alexnet	-	66.76	74.31	57.76	64.87	64.87	0.56	697.02
MobileNet_1_00	1.00	69.43	70.56	68.88	69.69	69.69	3.24	85.18
MobileNet_0_75	0.75	69.99	71.19	69.04	70.08	70.08	1.84	48.55
MobileNet_0_50	0.50	67.01	68.17	65.59	66.83	66.83	0.83	22.14
MobileNet_0_25	0.25	67.09	71.51	62.41	66.58	66.58	0.22	5.96
Our Model	-	**71.69**	**71.73**	**71.69**	**71.59**	**71.59**	**2.22**	**115.19**

จากตารางที่ 4.1 ผลการเปรียบเทียบโมเดลจำแนก 7 กลุ่ม พบว่า โมเดลที่สามารถจำแนกการแสดงอารมณ์ทางใบหน้าได้ดีที่สุด คือ โมเดล Our Model มีค่า Accuracy เท่ากับ 71.69 มีค่า precision เท่ากับ 71.73 มีค่า recall เท่ากับ 71.69 โดยมีการใช้พารามิเตอร์อยู่ที่ 2.22 ล้าน และ มี FLOPs อยู่ที่ 115.19 ล้าน

ตารางที่ 4.2 ผลความแม่นยำของโมเดล จำแนกอารมณ์ 7 กลุ่มอารมณ์

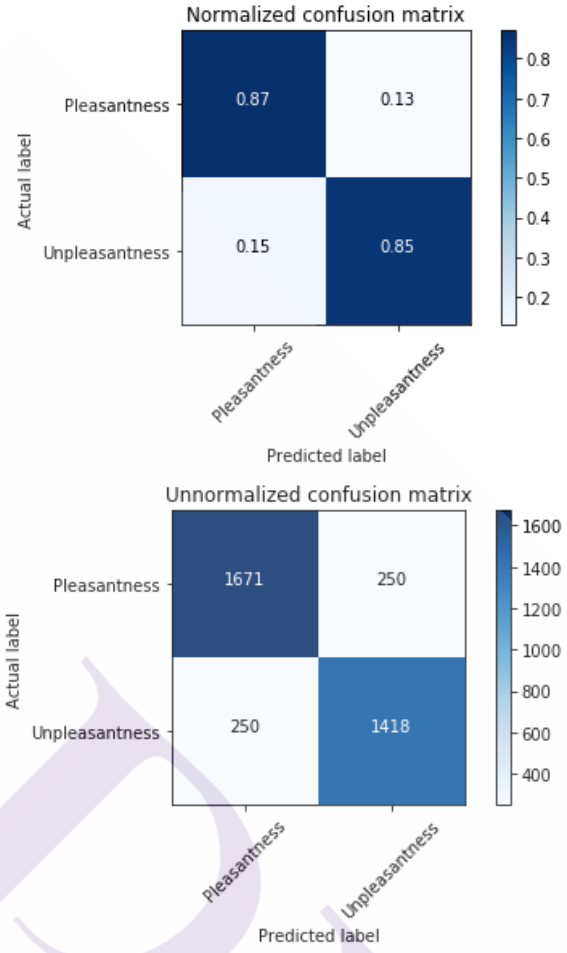
โมเดล จำแนกอารมณ์ 7 กลุ่มอารมณ์	กลุ่มที่บ่งบอกลักษณะว่ามีความสนใจ (Pleasantness)						กลุ่มที่บ่งบอกลักษณะว่าไม่มีความสนใจ (Unpleasantness)							
	มีความสุข (happy)		ประหลาดใจ (surprise)		ปกติ (neutral)		โกรธ (angry)		ไม่ชอบ (disgust)		กลัว (fear)		เศร้า (sad)	
	precision	recall	precision	recall	precision	recall	precision	recall	precision	recall	precision	recall	precision	recall
simple_CNN	89.19	80.16	79.57	71.34	67.57	54.51	50.51	59.76	38.18	56.76	25.38	55.37	55.89	49.04
simpler_CNN	89.76	83.05	75.48	72.35	71.57	55.86	51.73	57.60	50.91	62.22	28.98	50.83	51.35	49.51
tiny_XCEPTION	87.71	83.17	73.80	71.56	68.37	57.92	53.36	57.96	38.18	50.00	33.52	49.30	51.85	48.05
mini_XCEPTION	88.51	86.25	75.72	76.64	66.93	62.72	59.67	60.16	50.91	63.64	41.86	52.37	54.21	49.16
big_XCEPTION	85.32	85.13	77.88	76.78	69.81	61.72	52.95	65.16	65.45	66.67	49.43	54.38	55.89	51.47
tiny_Alexnet	88.05	85.34	79.81	74.61	75.40	56.94	55.19	63.62	60.00	75.00	35.42	64.26	55.05	50.54
MobileNet_1_00	86.01	87.10	81.97	81.38	71.57	62.14	59.67	62.34	69.09	90.48	54.92	58.12	54.88	57.19
MobileNet_0_75	85.89	89.45	81.97	83.37	71.57	61.88	62.93	63.19	70.91	84.78	54.36	63.50	56.06	53.28
MobileNet_0_50	83.50	88.01	82.93	81.37	63.26	63.06	57.64	57.76	70.91	82.98	51.52	55.51	56.57	49.70
MobileNet_0_25	85.89	85.50	77.64	81.36	69.81	59.78	60.49	60.86	69.09	67.86	48.30	53.46	51.01	54.40
Our Model	84.63	**88.96**	**85.44**	**84.62**	58.68	**71.25**	**61.88**	63.14	70.00	76.36	**71.22**	56.25	**66.80**	**57.91**

จากตารางที่ 4.2 ผลความแม่นยำของโมเดล จำแนกอารมณ์ 7 กลุ่มอารมณ์ พบว่า โมเดล Our Model สามารถจำแนกอารมณ์กลุ่ม ประหลาดใจ ได้ดีที่สุด เมื่อพิจารณาจากค่า precision เท่ากับ 85.44 และ recall เท่ากับ 84.62 รองลงมาคือ มีความสุข มีค่าค่า precision เท่ากับ 84.63 และ recall เท่ากับ 88.96 ตามลำดับ

4.1.2 ความแม่นยำของโมเดล จำแนกอารมณ์ 2 กลุ่มอารมณ์

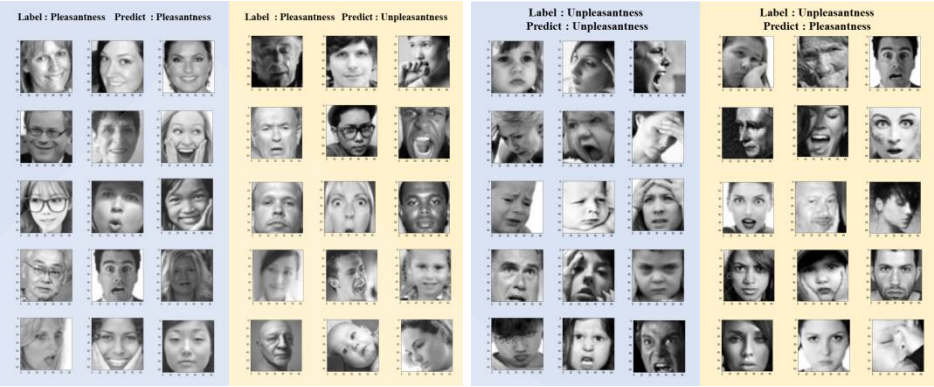
มีการจำแนกจำนวนกลุ่มไว้ทั้งหมด 2 กลุ่ม กลุ่มที่บ่งบอกลักษณะว่ามีความสนใจ และกลุ่มที่บ่งบอกลักษณะว่าไม่มีความสนใจ

พบว่า มีความถูกต้องในภาพรวม (Accuracy) คิดเป็นร้อยละ 86.07 กลุ่มที่บ่งบอกลักษณะว่ามีความสนใจ มีความถูกต้องมากที่สุด และรองลงมาคือ กลุ่มที่บ่งบอกลักษณะว่าไม่มีความสนใจ มีความถูกต้อง ตามลำดับ จึงสามารถสรุปได้ว่า ความแม่นยำของโมเดล จำแนกอารมณ์ 2 กลุ่มอารมณ์ มีความสามารถในการจำแนก กลุ่มที่บ่งบอกลักษณะว่ามีความสนใจ (Pleasantness) ได้ดี ดังภาพที่ 4.3



ภาพที่ 4.3 ผลความแม่นยำของโมเดล จำแนกอารมณ์ 2 กลุ่มอารมณ์

เมื่อพิจารณากลุ่มที่มีความถูกต้องน้อยที่สุดพบว่าเป็นกลุ่มที่บ่งบอกลักษณะว่าไม่มีความสนใจ แต่เมื่อพิจารณา ภาพที่ 4.3 ทำให้พบข้อผิดพลาดในการจำแนกเนื่องจาก กลุ่มที่บ่งบอกลักษณะว่ามีความสนใจ และกลุ่มที่บ่งบอกลักษณะว่าไม่มีความสนใจ มีการจำแนกผิดไป 250 จากชุดทดสอบข้อมูล ซึ่งมีจำนวนที่เท่ากัน ดังภาพที่ 4.3 มีลักษณะที่ใกล้เคียงและคล้ายคลึงกันมาก ซึ่งเป็นการยากที่จะจำแนกหรือแบ่งแยกทั้ง 2 กลุ่มนี้ได้ชัดเจน ดังภาพที่ 4.4



ภาพที่ 4.4 ภาพตัวอย่างกลุ่มที่มีความสนใจ (Pleasantness) และกลุ่มที่ไม่มีความสนใจ (Unpleasantness)

ตารางที่ 4.3 ผลการเปรียบเทียบโมเดลจำแนกอารมณ์ 2 กลุ่มอารมณ์

โมเดลจำแนกอารมณ์ 7 กลุ่มอารมณ์	alpha	acc	precision	recall	fbeta_score	fmeasure	Total params/M	Total FLOPS/M
simple_CNN	-	83.03	83.03	83.03	83.03	83.03	0.63	254.53
simpler_CNN	-	82.95	82.95	82.95	82.95	82.95	0.55	76.73
tiny_XCEPTION	-	81.86	81.86	81.86	81.86	81.86	0.02	6.72
mini_XCEPTION	-	84.03	84.03	84.03	84.03	84.03	0.05	20.14
big_XCEPTION	-	84.17	84.17	84.17	84.17	84.17	0.20	130.06
tiny_Alexnet	-	82.22	82.22	82.22	82.22	82.22	0.55	696.65
MobileNet_1_00	1.00	85.32	85.32	85.32	85.32	85.32	3.23	85.17
MobileNet_0_75	0.75	85.15	85.15	85.15	85.15	85.15	1.83	48.54
MobileNet_0_50	0.50	84.15	85.15	86.15	87.15	88.15	0.83	22.14
MobileNet_0_25	0.25	83.00	83.00	83.00	83.00	83.00	0.22	5.96
Our Model	-	**86.07**	**86.07**	**86.07**	**86.07**	**86.07**	**2.19**	**115.12**

จากตารางที่ 4.3 ผลการเปรียบเทียบโมเดลจำแนก 2 กลุ่ม พบว่า โมเดลที่สามารถจำแนกการแสดงอารมณ์ทางใบหน้าได้ดีที่สุด คือ โมเดล Our Model มีค่า Accuracy เท่ากับ 86.07 มีค่า precision เท่ากับ 86.07 มีค่า recall เท่ากับ 86.07 โดยมีการใช้พารามิเตอร์อยู่ที่ 2.19 ล้าน และ มี FLOPs อยู่ที่ 115.12 ล้าน

ตารางที่ 4.4 ผลความแม่นยำของโมเดล จำแนกอารมณ์ 2 กลุ่มอารมณ์

โมเดลจำแนกอารมณ์ 2 กลุ่มอารมณ์	กลุ่มที่บ่งบอกลักษณะว่ามีความสนใจ (Pleasantness)		กลุ่มที่บ่งบอกลักษณะว่าไม่มีความสนใจ (Unpleasantness)	
	precision	recall	precision	recall
simple_CNN	83.91	84.31	82.01	81.57
simpler_CNN	82.40	85.25	83.57	80.48
tiny_XCEPTION	82.77	83.25	80.82	80.29
mini_XCEPTION	82.56	86.95	85.73	81.02
big_XCEPTION	83.24	86.67	85.25	81.54
tiny_Alexnet	84.70	82.55	79.38	81.83
MobileNet_1_00	85.74	86.68	84.83	83.78
MobileNet_0_75	82.87	88.64	87.77	81.65
MobileNet_0_50	85.06	85.28	83.09	82.85
MobileNet_0_25	83.86	84.30	82.01	81.53
Our Model	**86.99**	**86.99**	**85.01**	**85.01**

จากตารางที่ 4.4 ผลความแม่นยำของโมเดล จำแนกอารมณ์ 2 กลุ่มอารมณ์ พบว่าโมเดล Our Model สามารถจำแนกอารมณ์กลุ่ม กลุ่มที่บ่งบอกลักษณะว่ามีความสนใจ ได้ดีที่สุด เมื่อพิจารณาจากค่า precision เท่ากับ 86.99 และ recall เท่ากับ 86.99 รองลงมาคือ มีกลุ่มที่บ่งบอกลักษณะว่าไม่มีความสนใจ มีค่าค่า precision เท่ากับ 85.01 และ recall เท่ากับ 85.01 ตามลำดับ

4.2 เวลา/ทรัพยากรที่ใช้ในการจำแนกอารมณ์บนอุปกรณ์ฝังตัว(Raspberry Pi)

Dataset 1. คลิปวีดีโอ 15วิ 30 fps 1280x720 โดยการ Sampling ภาพ 1 วินาที 1 ภาพ และ 2 วินาที 1 ภาพ เพื่อนำมาใช้ทดสอบเวลา/ทรัพยากรที่ใช้ในการจำแนกอารมณ์บนอุปกรณ์ฝังตัว (Raspberry Pi)

ขั้นตอนการทำงานบน Raspberry Pi

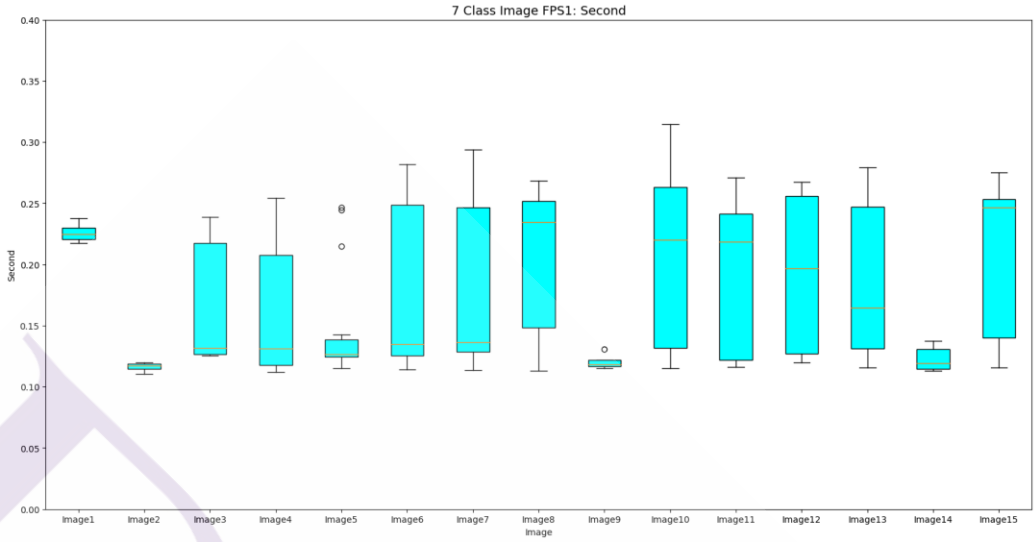
1. Sampling ภาพจากวีดีโอ (1 วิ 1 ภาพ/ 2 วิ 1 ภาพ)
2. ทำ face detection จากภาพ ในข้อที่ 1 โดยใช้ Opencv
3. ส่งแต่ละ face มาจำแนกอารมณ์โมเดลที่ปรับปรุงขึ้นมา พร้อมจับเวลา เฉพาะตอนจำแนกอารมณ์
4. ทำซ้ำข้อ 1 – 3 จนกว่าจะจบวีดีโอ
5. ทำซ้ำข้อ 1 – 4 จำนวน 4 รอบ

4.2.1 เวลา/ทรัพยากรที่ใช้ของโมเดล จำแนกอารมณ์ 7 กลุ่มอารมณ์

ตารางที่ 4.5 ผลการทดสอบการจำแนกใบหน้าบนอุปกรณ์ฝังตัว (Raspberry Pi) สำหรับโมเดล 7 กลุ่ม โดยสุ่มภาพ 1 วินาที ต่อ 1 ภาพ

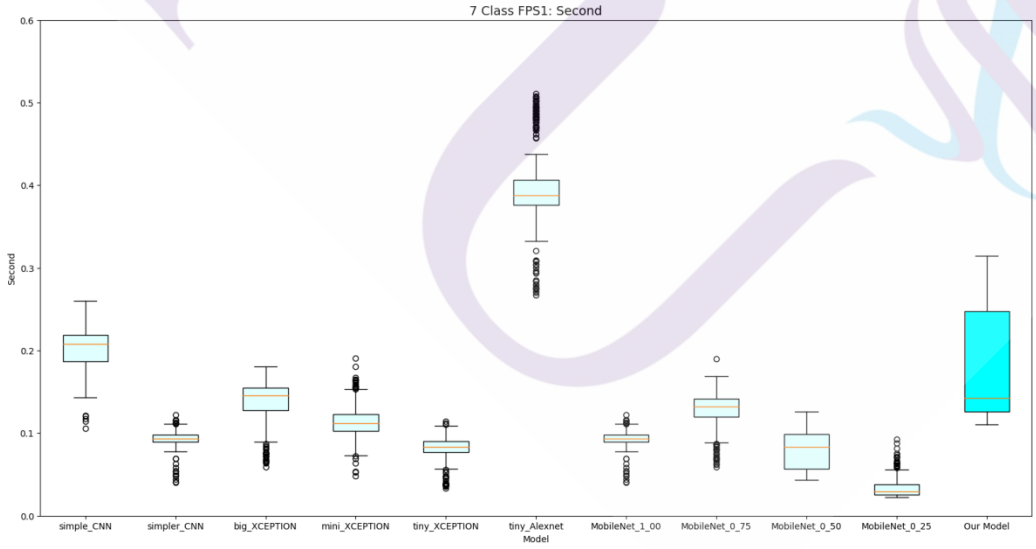
โมเดลจำแนกอารมณ์ 7 กลุ่มอารมณ์	FPS1: Second	CPU temp	CPU usage	RAM total	RAM used	Total params/M	Total FLOPS/M
simple_CNN	0.20	66.06	77.37	896.70	323.79	0.64	254.90
simpler_CNN	0.09	60.76	0.75	896.70	322.67	0.56	76.76
big_XCEPTION	0.14	54.22	0.78	896.70	360.63	0.02	6.81
mini_XCEPTION	0.12	53.64	0.78	896.70	373.71	0.06	20.33
tiny_XCEPTION	0.08	56.40	0.76	896.70	380.88	0.21	131.54
tiny_Alexnet	0.40	57.73	0.82	896.70	351.33	0.56	697.02
MobileNet_1_00	0.09	60.76	0.75	896.70	322.65	3.24	85.18
MobileNet_0_75	0.14	57.64	0.75	896.70	367.28	1.84	48.55
MobileNet_0_50	0.08	61.38	0.75	896.70	366.47	0.83	22.14
MobileNet_0_25	0.03	62.86	0.75	896.70	364.81	0.22	5.96
Our Model	**0.18**	**56.90**	**77.79**	**896.70**	**344.01**	**2.22**	**115.19**

จากตารางที่ 4.5 ผลการทดสอบการจำแนกใบหน้าบนอุปกรณ์ฝังตัว (Raspberry Pi) พบว่า ความเร็วในการจำแนกใบหน้าบนอุปกรณ์ฝังตัว (Raspberry Pi) สำหรับโมเดล 7 กลุ่ม โดยสุ่มภาพ 1 วินาที ต่อ 1 ภาพ ซึ่งโมเดล Our Model มีความเร็วอยู่ที่ 0.18 วินาที CPU มีอุณหภูมิอยู่ที่ 56.90 องศาเซลเซียล มีการใช้งาน CPU คิดเป็นร้อยละ 77.79 ใช้งาน Ram ได้จำนวน 344.01 Mb มีจำนวน Total params/M เท่ากับ 2.22 และจำนวน Total FLOPS/M เท่ากับ 115.19 ดังภาพที่ 4.5 และ 4.6



ภาพที่ 4.5 เวลาที่ใช้ทดสอบการจำแนกใบหน้าบนอุปกรณ์ฝังตัว (Raspberry Pi) สำหรับโมเดล 7 กลุ่ม โดยสุ่มภาพ 1 วินาที ต่อ 1 ภาพ

จากภาพที่ 4.5 จากการทดสอบจำแนกอารมณ์บนอุปกรณ์ฝังตัว (Raspberry Pi) กันวีดีโอ 15 วินาที จำนวน 4 รอบ พบว่า เมื่อตรวจจับจำนวน 4 รอบได้เท่ากับ 28 ใบหน้า มีภาพที่ 8 และ 13 โดยมีความเร็งเฉลี่ย 0.203361222 , 0.189860593 ตามลำดับ

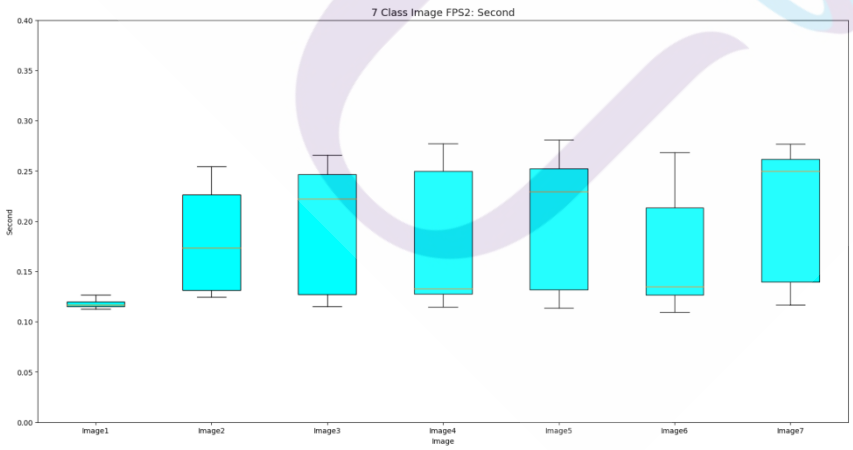


ภาพที่ 4.6 ผลการเปรียบเทียบเวลาสำหรับโมเดลจำแนก 7 กลุ่มอารมณ์ โดยสุ่มภาพ 1 วินาที ต่อ 1 ภาพ

ตารางที่ 4.6 ผลการทดสอบการจำแนกใบหน้าบนอุปกรณ์ฝังตัว (Raspberry Pi) สำหรับโมเดล 7 กลุ่ม โดยสุ่มภาพ 2 วินาที ต่อ 1 ภาพ

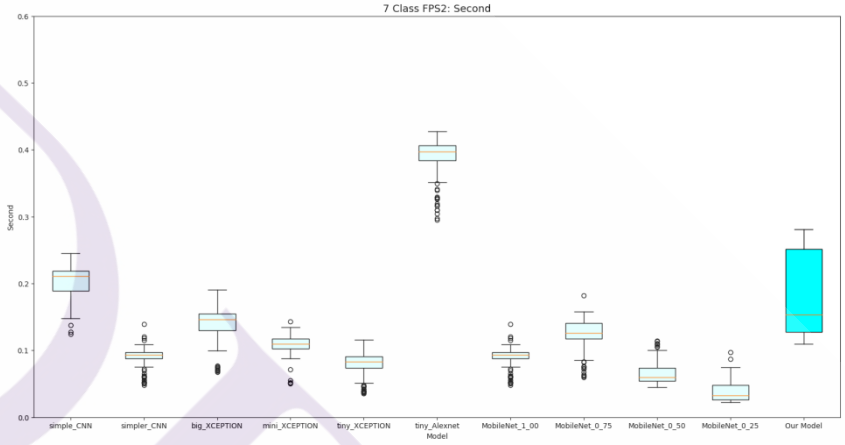
โมเดลจำแนกอารมณ์ 7 กลุ่มอารมณ์	FPS1: Second	CPU temp	CPU usage	RAM total	RAM used	Total params/M	Total FLOPS/M
simple_CNN	0.20	56.54	0.78	896.70	375.21	0.64	254.90
simpler_CNN	0.09	61.21	0.76	896.70	323.48	0.56	76.76
big_XCEPTION	0.14	55.73	0.78	896.70	361.55	0.02	6.81
mini_XCEPTION	0.11	53.96	0.77	896.70	376.31	0.06	20.33
tiny_XCEPTION	0.08	58.08	0.75	896.70	381.19	0.21	131.54
tiny_Alexnet	0.39	60.06	0.79	896.70	314.37	0.56	697.02
MobileNet_1_00	0.09	61.21	0.76	896.70	323.44	3.24	85.18
MobileNet_0_75	0.12	58.83	0.74	896.70	369.03	1.84	48.55
MobileNet_0_50	0.07	62.79	0.73	896.70	365.86	0.83	22.14
MobileNet_0_25	0.04	62.05	0.74	896.70	365.29	0.22	5.96
Our Model	0.19	57.42	0.76	896.70	346.25	2.22	115.19

จากตารางที่ 4.6 ผลการทดสอบการจำแนกใบหน้าบนอุปกรณ์ฝังตัว (Raspberry Pi) พบว่า ความเร็วในการจำแนกใบหน้าบนอุปกรณ์ฝังตัว (Raspberry Pi) สำหรับโมเดล 7 กลุ่ม โดยสุ่มภาพ 2 วินาที ต่อ 1 ภาพ ซึ่งโมเดล Our Model มีความเร็วอยู่ที่ 0.19 วินาที CPU มีอุณหภูมิอยู่ที่ 57.42 องศาเซลเซียล มีการใช้งาน CPU คิดเป็นร้อยละ 76 ใช้งาน Ram ได้จำนวน 346.25 Mb มีจำนวน Total params/M เท่ากับ 2.22 และจำนวน Total FLOPS/M เท่ากับ 115.19 ดังภาพที่ 4.7 และ 4.8

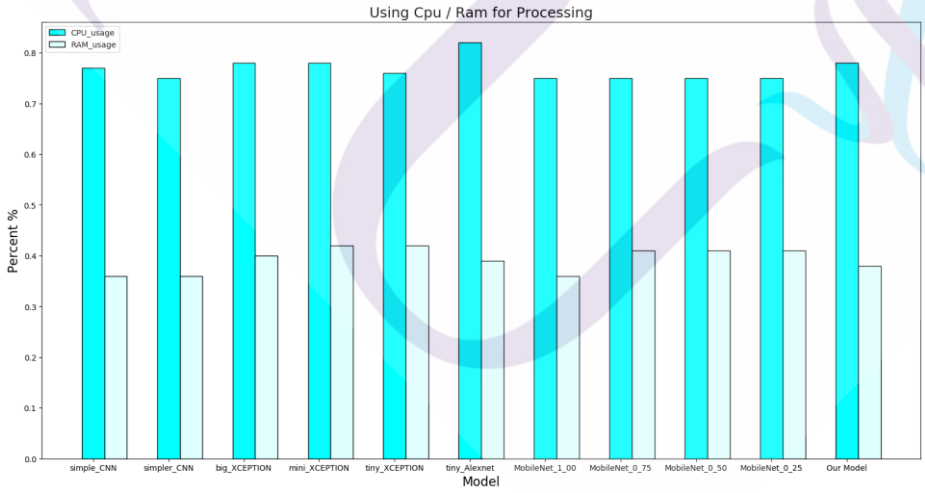


ภาพที่ 4.7 เวลาที่ใช้ทดสอบการจำแนกใบหน้าบนอุปกรณ์ฝังตัว (Raspberry Pi) สำหรับโมเดล 7 กลุ่ม โดยสุ่มภาพ 2 วินาที ต่อ 1 ภาพ

จากภาพที่ 4.7 จากการทดสอบจำแนกอารมณ์บนอุปกรณ์ฝังตัว (Raspberry Pi) กับวีดีโอ 15 วินาที จำนวน 4 รอบ โดยใช้ เวลา 2 วินาที ต่อ ภาพ 1 ภาพ พบว่า เมื่อตรวจจับจำนวน 4 รอบได้เท่ากับ 28 ใบหน้า มีภาพที่ 4 และ 5 โดยมีความเร็งเฉลี่ย 0.170171185 , 0.196908037 ตามลำดับ



ภาพที่ 4.8 ผลการเปรียบเทียบเวลาสำหรับโมเดลจำแนก 7 กลุ่มอารมณ์ โดยสุ่มภาพ 2 วินาที ต่อ 1 ภาพ



ภาพที่ 4.9 ผลการเปรียบเทียบการใช้งาน CPU และ Ram บนอุปกรณ์ฝังตัว (Raspberry Pi) ของโมเดล Our Model

จากภาพที่ 4.9 ผลการเปรียบเทียบการใช้งาน CPU และ Ram บนอุปกรณ์ฝังตัว (Raspberry Pi) ของ Our Model พบว่า การใช้เวลา/ทรัพยากรที่ใช้ในการจำแนก 7 กลุ่มอารมณ์บนอุปกรณ์ฝังตัว (Raspberry Pi) ของโมเดล Our Model มีการใช้งานที่ใกล้เคียงกับโมเดลอื่นๆ

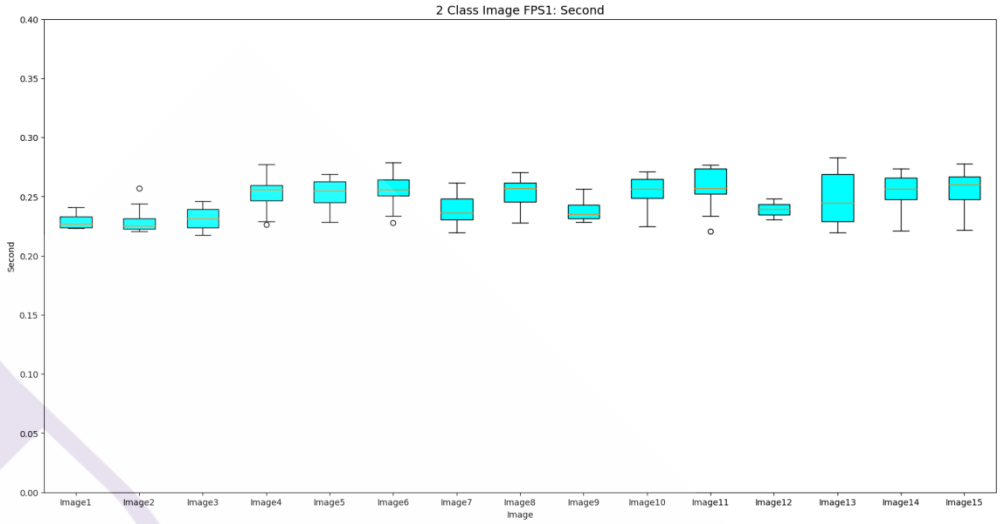
4.2.2 เวลา/ทรัพยากรที่ใช้ของโมเดล จำแนกอารมณ์ 2 กลุ่มอารมณ์

ตารางที่ 4.7 ผลการทดสอบการจำแนกใบหน้าบนอุปกรณ์ฝังตัว (Raspberry Pi) สำหรับโมเดล 2 กลุ่มโดยสุ่มภาพ 1 วินาที ต่อ 1 ภาพ

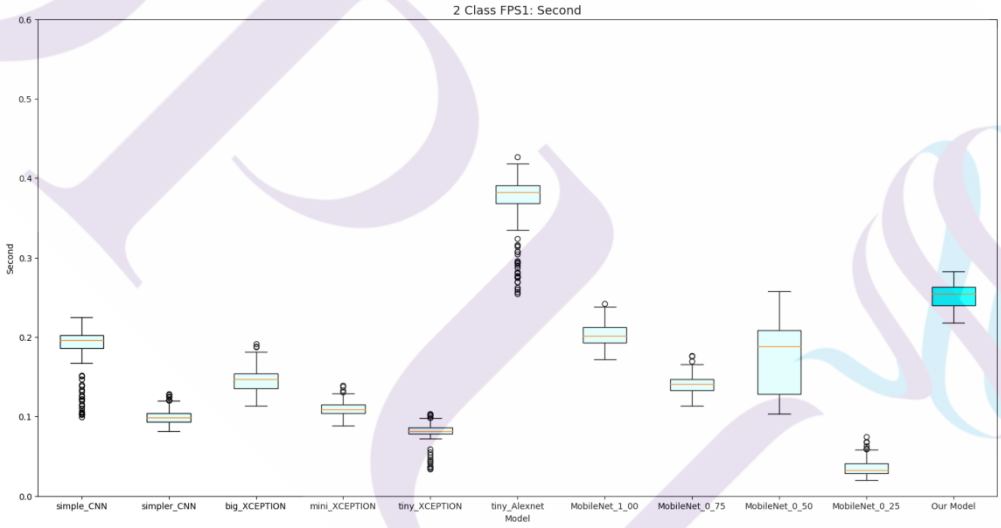
โมเดลจำแนกอารมณ์ 2 กลุ่มอารมณ์	FPS1: Second	CPU temp	CPU usage	RAM total	RAM used	Total params/M	Total FLOPS/M
simple_CNN	0.19	54.47	0.79	896.70	408.76	0.63	254.53
simpler_CNN	0.10	55.95	0.77	896.70	453.61	0.55	76.73
big_XCEPTION	0.14	53.27	0.77	896.70	391.30	0.02	6.72
mini_XCEPTION	0.11	53.71	0.77	896.70	405.27	0.05	20.14
tiny_XCEPTION	0.08	53.85	0.77	896.70	416.44	0.20	130.06
tiny_Alexnet	0.37	56.92	0.81	896.70	452.14	0.55	696.65
MobileNet_1_00	0.20	53.36	0.73	896.70	545.72	3.23	85.17
MobileNet_0_75	0.14	55.39	0.75	896.70	503.01	1.83	48.54

โมเดลจำแนกอารมณ์ 2 กลุ่มอารมณ์	FPS1: Second	CPU temp	CPU usage	RAM total	RAM used	Total params/M	Total FLOPS/M
MobileNet_0_50	0.17	58.05	0.72	896.70	490.42	0.83	22.14
MobileNet_0_25	0.04	61.64	0.74	896.70	448.25	0.22	5.96
Our Model	0.25	50.52	0.77	896.70	423.68	2.19	115.12

จากตารางที่ 4.7 ผลการทดสอบการจำแนกใบหน้าบนอุปกรณ์ฝังตัว (Raspberry Pi) พบว่า ความเร็วในการจำแนกใบหน้าบนอุปกรณ์ฝังตัว (Raspberry Pi) สำหรับโมเดล 2 กลุ่ม โดยสุ่มภาพ 1 วินาที ต่อ 1 ภาพ ซึ่งโมเดล Our Model มีความเร็วอยู่ที่ 0.25 วินาที CPU มีอุณหภูมิอยู่ที่ 50.52 องศาเซลเซียล มีการใช้งาน CPU คิดเป็นร้อยละ 77 ใช้งาน Ram ได้จำนวน 423.68 Mb มีจำนวน Total params/M เท่ากับ 2.19 และจำนวน Total FLOPS/M เท่ากับ 115.12 ดังภาพที่ 4.10 และ 4.11



ภาพที่ 4.10 เวลาที่ใช้ทดสอบการจำแนกใบหน้าบนอุปกรณ์ฝังตัว (Raspberry Pi) สำหรับโมเดล 2 กลุ่ม โดยสุ่มภาพ 1 วินาที ต่อ 1 ภาพ

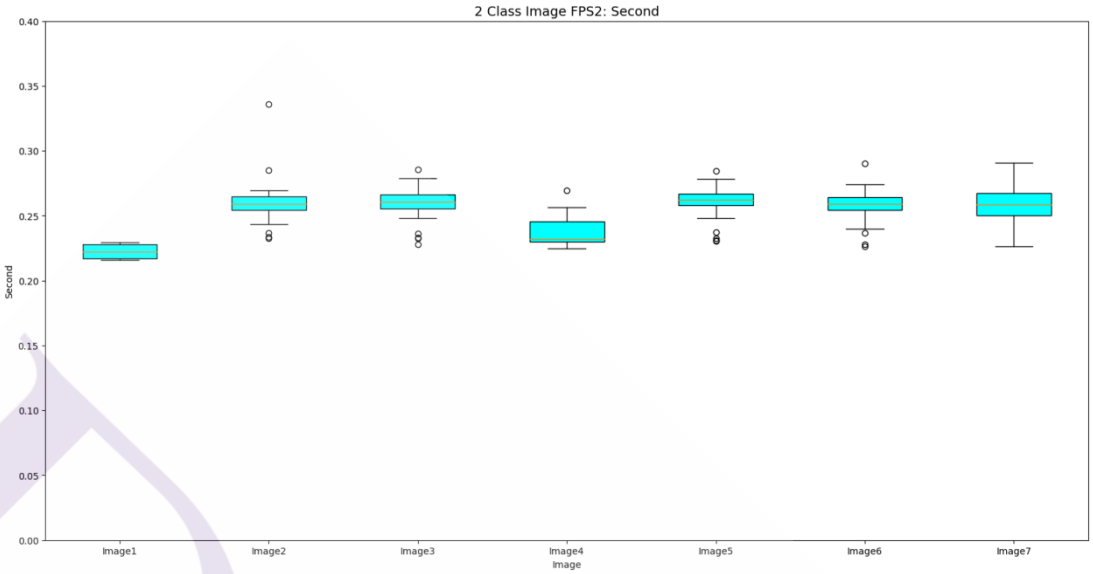


ภาพที่ 4.11 ผลการเปรียบเทียบเวลาสำหรับโมเดลจำแนก 2 กลุ่มอารมณ์ โดยสุ่มภาพ 1 วินาที ต่อ 1 ภาพ

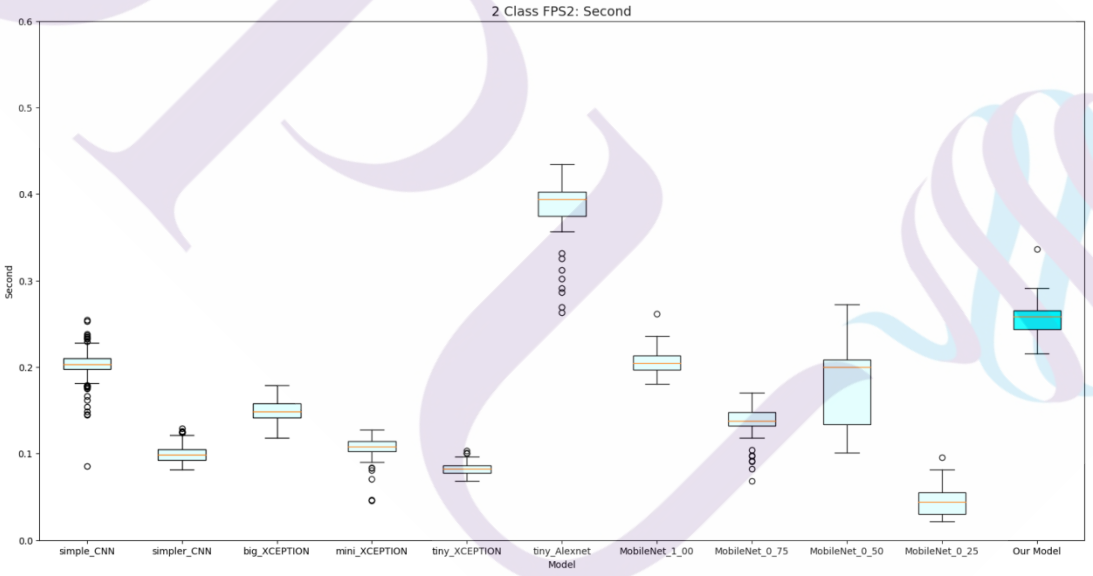
ตารางที่ 4.8 ผลการทดสอบการจำแนกใบหน้าบนอุปกรณ์ฝังตัว (Raspberry Pi) สำหรับโมเดล 2 กลุ่ม สุ่มภาพ 2 วินาที ต่อ 1 ภาพ

โมเดลจำแนกอารมณ์ 2 กลุ่มอารมณ์	FPS1: Second	CPU temp	CPU usage	RAM total	RAM used	Total params/M	Total FLOPS/M
simple_CNN	0.20	54.38	0.78	896.70	408.82	0.63	254.53
simpler_CNN	0.10	53.07	0.76	896.70	453.54	0.55	76.73
big_XCEPTION	0.15	53.75	0.78	896.70	392.78	0.02	6.72
mini_XCEPTION	0.11	53.17	0.76	896.70	404.73	0.05	20.14
tiny_XCEPTION	0.08	53.79	0.76	896.70	415.85	0.20	130.06
tiny_Alexnet	0.39	56.20	0.81	896.70	450.84	0.55	696.65
MobileNet_1_00	0.21	53.10	0.71	896.70	531.48	3.23	85.17
MobileNet_0_75	0.14	54.85	0.73	896.70	486.97	1.83	48.54
MobileNet_0_50	0.18	58.90	0.72	896.70	487.62	0.83	22.14
MobileNet_0_25	0.04	58.78	0.74	896.70	447.67	0.22	5.96
Our Model	0.26	51.05	0.75	896.70	424.58	2.19	115.12

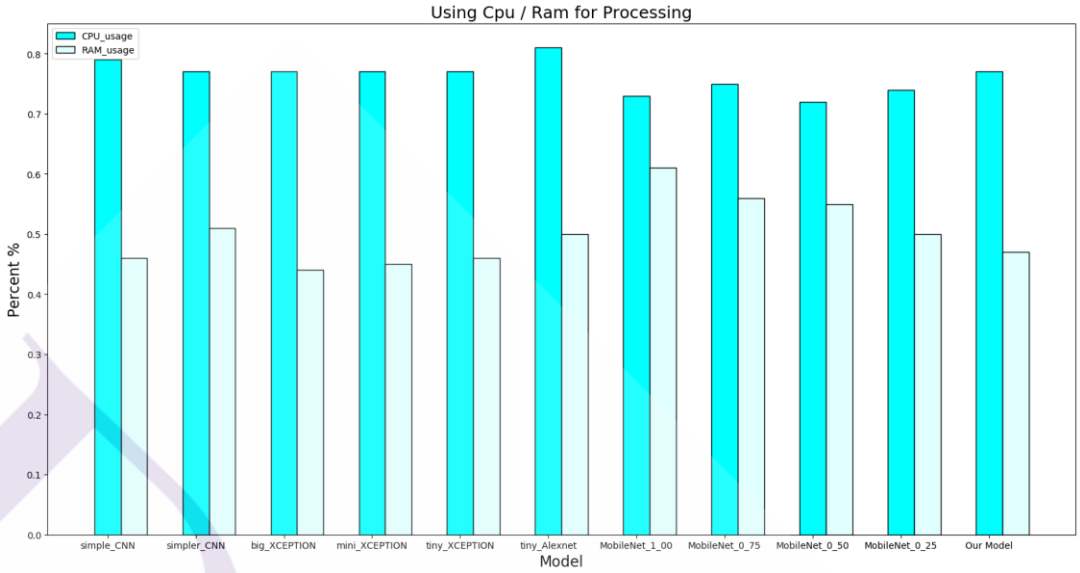
จากตารางที่ 4.8 ผลการทดสอบการจำแนกใบหน้าบนอุปกรณ์ฝังตัว (Raspberry Pi) พบว่า ความเร็วในการจำแนกใบหน้าบนอุปกรณ์ฝังตัว (Raspberry Pi) สำหรับโมเดล 2 กลุ่ม โดยสุ่มภาพ 2 วินาที ต่อ 1 ภาพ ซึ่งโมเดล Our Model มีความเร็วอยู่ที่ 0. 26 วินาที CPU มีอุณหภูมิอยู่ที่ 51.05 องศาเซลเซียล มีการใช้งาน CPU คิดเป็นร้อยละ 77 ใช้งาน Ram ได้จำนวน 424.58 Mb มีจำนวน Total params/M เท่ากับ 2.19 และจำนวน Total FLOPS/M เท่ากับ 115.12 ดังภาพที่ 4.12 และ 4.13



ภาพที่ 4.12 เวลาที่ใช้ทดสอบการจำแนกใบหน้าบนอุปกรณ์ฝังตัว (Raspberry Pi) สำหรับโมเดล 2 กลุ่ม โดยสุ่มภาพ 2 วินาที ต่อ 1 ภาพ



ภาพที่ 4.13 ผลการเปรียบเทียบเวลาสำหรับโมเดลจำแนก 2 กลุ่มอารมณ์ โดยสุ่มภาพ 2 วินาที ต่อ 1 ภาพ



ภาพที่ 4.14 ผลการเปรียบเทียบการใช้งาน CPU และ Ram บนอุปกรณ์ฝังตัว (Raspberry Pi) ของ Our Model

จากภาพที่ 4.14 ผลการเปรียบเทียบการใช้งาน CPU และ Ram บนอุปกรณ์ฝังตัว (Raspberry Pi) ของ Our Model พบว่า การใช้เวลา/ทรัพยากรที่ใช้ในการจำแนก 2 กลุ่มอารมณ์บนอุปกรณ์ฝังตัว (Raspberry Pi) ของโมเดล Our Model มีการใช้งานที่ใกล้เคียงกับโมเดลอื่นๆ

บทที่ 5
บทสรุปและข้อเสนอแนะ

วิจัยเรื่องนี้มีวัตถุประสงค์เพื่อจำแนกอารมณ์จากใบหน้าแบบ Real-time โดยใช้โครงสร้างเครือข่ายประสาทเทียมแบบคอนโวลูชั่น ที่มีประสิทธิภาพเทียบเคียงกับโมเดลซับซ้อนที่มีความแม่นยำในการทำนายสูง และโมเดลที่ได้มีขนาดเล็กพอที่จะสามารถใช้งานบน Raspberry Pi ได้ โดยสามารถ สรุป อภิปราย และข้อเสนอแนะงานวิจัยได้ดังนี้

5.1 สรุปผลการวิจัย

5.1.1 การวัดความแม่นยำของโมเดลบนเครื่องคอมพิวเตอร์สมรรถนะสูง

5.1.1.1 ความแม่นยำของโมเดล จำแนกอารมณ์ 7 กลุ่มอารมณ์

สรุปผลความแม่นยำของโมเดล จำแนกอารมณ์ 7 กลุ่มอารมณ์ มีการจำแนกจำนวนกลุ่มไว้ทั้งหมด 7 กลุ่ม ได้แก่ ใบหน้าโกรธ, ใบหน้าไม่ชอบ, ใบหน้ากลัว, ใบหน้ามีความสุข, ใบหน้าเศร้า, ใบหน้าประหลาดใจ และใบหน้าแบบปกติ พบว่า มีความถูกต้องในภาพรวม (Accuracy) คิดเป็นร้อยละ 71.69 สามารถจำแนกลักษณะการแสดงอารมณ์จากใบหน้ามีความสุข มีความถูกต้องมากที่สุด และพบว่า กลุ่มที่มีความถูกต้องน้อยที่สุดพบว่าเป็นกลุ่มของใบหน้ากลัว และใบหน้าเศร้า ที่มีความถูกต้องเพียง 57-60% เท่านั้น จากการสังเกตภาพต้นฉบับที่ผ่านการจำแนกทำให้สามารถอธิบายได้ว่ากลุ่มของใบหน้ากลัว และใบหน้าเศร้า มีลักษณะที่ใกล้เคียงและคล้ายคลึงกันมาก ซึ่งเป็นการยากที่จะจำแนกหรือแบ่งแยกทั้ง 2 กลุ่มนี้ได้ชัดเจน โดยมีผลความแม่นยำของโมเดลล Our Model ดังต่อไปนี้ จำแนกอารมณ์กลุ่ม ประหลาดใจ ได้ดีที่สุด เมื่อพิจารณาจากค่า precision เท่ากับ 85.44 และ recall เท่ากับ 84.62 รองลงมาคือ มีความสุข มีค่าค่า precision เท่ากับ 84.63 และ recall เท่ากับ 88.96 ตามลำดับ และมีการใช้พารามิเตอร์อยู่ที่ 2.22 ล้าน และ มี FLOPs อยู่ที่ 115.19 ล้าน

5.1.1.2 ความแม่นยำของโมเดล จำแนกอารมณ์ 2 กลุ่มอารมณ์

สรุปผลความแม่นยำของโมเดล จำแนกอารมณ์ 2 กลุ่มอารมณ์ มีการจำแนกจำนวนกลุ่มไว้ทั้งหมด 2 กลุ่ม กลุ่มที่บ่งบอกลักษณะว่ามีความสนใจ และกลุ่มที่บ่งบอกลักษณะว่าไม่มีความสนใจ พบว่า มีความถูกต้องในภาพรวมคิดเป็นร้อยละ 86.07 กลุ่มที่บ่งบอกลักษณะว่ามีความสนใจ

มีความถูกต้องมากที่สุด และรองลงมาคือ กลุ่มที่บ่งบอกลักษณะว่าไม่มีความสนใจ มีความถูกต้องตามลำดับ จึงสามารถสรุปได้ว่า ความแม่นยำของโมเดล จำแนกอารมณ์ 2 กลุ่มอารมณ์ มีความสามารถในการจำแนก กลุ่มที่บ่งบอกลักษณะว่ามีความสนใจ ได้ดี เมื่อพิจารณากลุ่มที่มีความถูกต้องน้อยที่สุดพบว่าเป็นกลุ่มที่บ่งบอกลักษณะว่าไม่มีความสนใจ แต่เมื่อพิจารณา ภาพที่ 4.3 ทำให้พบข้อผิดพลาดในการจำแนกเนื่องจาก กลุ่มที่บ่งบอกลักษณะว่ามีความสนใจ และกลุ่มที่บ่งบอกลักษณะว่าไม่มีความสนใจ มีการจำแนกผิดไป 250 จากชุดทดสอบข้อมูล ซึ่งมีจำนวนที่เท่ากัน ดังภาพที่ 4.3 มีลักษณะที่ใกล้เคียงและคล้ายคลึงกันมาก ซึ่งเป็นการยากที่จะจำแนกหรือแบ่งแยกทั้ง 2 กลุ่มนี้ได้ชัดเจน มีการใช้พารามิเตอร์อยู่ที่ 2.19 ล้าน และ มี FLOPs อยู่ที่ 115.12 ล้าน

5.1.2. เวลา/ทรัพยากรที่ใช้ในการจำแนกอารมณ์บนอุปกรณ์ฝังตัว (Raspberry Pi)

5.1.2.1 เวลา/ทรัพยากรที่ใช้ของโมเดล จำแนกอารมณ์ 7 กลุ่มอารมณ์

สรุปผลการทดสอบการจำแนกใบหน้าบนอุปกรณ์ฝังตัวพบว่า ความเร็วในการจำแนกใบหน้าบนอุปกรณ์ฝังตัวสำหรับโมเดล 7 กลุ่ม โดยใช้ 1 วินาที ต่อ 1 ภาพ ซึ่งโมเดลที่นำเสนอมีความเร็วอยู่ที่ 0.18 วินาที CPU มีอุณหภูมิอยู่ที่ 56.90 องศาเซลเซียล มีการใช้งาน CPU คิดเป็นร้อยละ 77.79 ใช้งาน Ram ได้จำนวน 344.01 Mb มีจำนวน Total params/M เท่ากับ 2.22 และจำนวน Total FLOPS/M เท่ากับ 115.19

สรุปผลการทดสอบการจำแนกใบหน้าบนอุปกรณ์ฝังตัวพบว่า ความเร็วในการจำแนกใบหน้าบนอุปกรณ์ฝังตัวสำหรับโมเดล 7 กลุ่ม โดยใช้ 2 วินาที ต่อ 1 ภาพ ซึ่งโมเดลที่นำเสนอมีความเร็วอยู่ที่ 0.19 วินาที CPU มีอุณหภูมิอยู่ที่ 57.42 องศาเซลเซียล มีการใช้งาน CPU คิดเป็นร้อยละ 76 ใช้งาน Ram ได้จำนวน 346.25 Mb มีจำนวน Total params/M เท่ากับ 2.22 และจำนวน Total FLOPS/M เท่ากับ 115.19

เมื่อพิจารณา ความเร็วในการจำแนกใบหน้าบนอุปกรณ์ฝังตัวสำหรับโมเดล 7 กลุ่ม โดยใช้ 1 วินาที ต่อ 1 ภาพ และ 2 วินาที ต่อ 1 ภาพ พบว่า เวลา/ทรัพยากรในการจำแนกใบหน้าบนอุปกรณ์ฝังตัวสำหรับโมเดล 7 กลุ่ม โดยเวลาเฉลี่ยน ใช้เวลา 1 วินาที ต่อ 1 ภาพ ให้ค่าเฉลี่ยที่ดีกว่า

5.1.3 เวลา/ทรัพยากรที่ใช้ของโมเดล จำแนกอารมณ์ 2 กลุ่มอารมณ์

สรุปผลการทดสอบการจำแนกใบหน้าบนอุปกรณ์ฝังตัวพบว่า ความเร็วในการจำแนกใบหน้าบนอุปกรณ์ฝังตัวสำหรับโมเดล 2 กลุ่ม โดยใช้ 1 วินาที ต่อ 1 ภาพ ซึ่งโมเดลที่นำเสนอมีความเร็วอยู่ที่ 0.25 วินาที CPU มีอุณหภูมิอยู่ที่ 50.52 องศาเซลเซียล มีการใช้งาน CPU คิดเป็นร้อยละ 77 ใช้งาน Ram ได้จำนวน 423.68 Mb มีจำนวน Total params/M เท่ากับ 2.19 และจำนวน Total FLOPS/M เท่ากับ 115.12

สรุปผลการทดสอบการจำแนกใบหน้าบนอุปกรณ์ฝังตัวพบว่า ความเร็วในการจำแนกใบหน้าบนอุปกรณ์ฝังตัว (Raspberry Pi) สำหรับโมเดล 2 กลุ่ม โดยใช้ 2 วินาที ต่อ 1 ภาพ ซึ่งโมเดลที่นำเสนอมีความเร็วอยู่ที่ 0. 26 วินาที CPU มีอุณหภูมิอยู่ที่ 51.05 องศาเซลเซียล มีการใช้งาน CPU คิดเป็นร้อยละ 77 ใช้งาน Ram ได้จำนวน 424.58 Mb มีจำนวน Total params/M เท่ากับ 2.19 และจำนวน Total FLOPS/M เท่ากับ 115.12

เมื่อพิจารณา ความเร็วในการจำแนกใบหน้าบนอุปกรณ์ฝังตัวสำหรับโมเดล 2กลุ่ม โดยใช้ 1 วินาที ต่อ 1 ภาพ และ 2 วินาที ต่อ 1 ภาพ พบว่า เวลา/ทรัพยากรในการจำแนกใบหน้าบนอุปกรณ์ฝังตัวสำหรับโมเดล 7 กลุ่ม โดยเวลาเฉลี่ยน ใช้เวลา 1 วินาที ต่อ 1 ภาพ ให้ค่าเฉลี่ยที่ดีกว่า

สรุปผลจากการทดสอบการจำแนกใบหน้าบนอุปกรณ์ฝังตัวสำหรับโมเดล 7 กลุ่ม และ 2 กลุ่ม โดยใช้เวลา 1 วินาที ต่อ 1 ภาพ และใช้เวลา 2 วินาที ต่อ 1 ภาพ ความเร็วในการประมวลผลใช้ใกล้เคียงกัน เนื่องจากใช้เวลาในการจำแนกไม่แตกต่างกันและใช้เวลาน้อยกว่า 1 วินาที

5.2 อภิปรายผลการวิจัย

การจำแนกอารมณ์จากใบหน้าแบบ Real-time โดยใช้ โครงสร้างเครือข่ายประสาทเทียมแบบคอนโวลูชั่น ซึ่งสอดคล้องกับ Octavio Arriaga and Paul G. Ploger (2017) โดยสร้างการตรวจหาใบหน้าแบบเรียลไทม์ พร้อมกับการจำแนกเพศและการจำแนกอารมณ์ โดยใช้เครือข่ายประสาทเทียมแบบคอนโวลูชั่นเพื่อใช้งานในหุ่นยนต์ Care-O-bot 3 ซึ่งสามารถจำแนกได้ 7 กลุ่มอารมณ์ นอกจากนี้โมเดลที่ได้มีประสิทธิภาพเทียบเคียงกับโมเดล state-of-art: Xception Model ที่มีการจำแนกอารมณ์ 7 ได้กลุ่ม พบว่า มีความถูกต้องในภาพรวม คิดเป็นร้อยละ 71.69 ซึ่งกลุ่มที่มีความถูกต้องน้อยที่สุดเป็นกลุ่มของใบหน้ากลัว และใบหน้าเศร้า ที่มีความถูกต้องเพียง 57-60% เท่านั้น จากการสังเกตภาพต้นฉบับที่ผ่านการจำแนกทำให้สามารถอธิบายได้ว่ากลุ่มของใบหน้ากลัว (fear) และใบหน้าเศร้า มีลักษณะที่ใกล้เคียงและคล้ายคลึงกันมาก ซึ่งเป็นการยากที่จะจำแนกหรือแบ่งแยกทั้ง 2 กลุ่มนี้ได้ชัดเจน ต่อมาโมเดลที่มีการจำแนกอารมณ์เป็น 2 กลุ่ม คือ มีความสนใจ และไม่มีความสนใจ ซึ่งกลุ่มที่มีความถูกต้องน้อยที่สุดพบว่าเป็นกลุ่มที่บ่งบอกลักษณะว่าไม่มีความสนใจ แต่เมื่อพิจารณาและทำให้พบข้อผิดพลาดในการจำแนกเนื่องจาก กลุ่มที่บ่งบอกลักษณะว่ามีความสนใจ และกลุ่มที่บ่งบอกลักษณะว่าไม่มีความสนใจ มีการจำแนกผิดไป 250 จากชุดทดสอบข้อมูล ซึ่งมีจำนวนที่เท่ากัน มีลักษณะที่ใกล้เคียงและคล้ายคลึงกันมาก ซึ่งเป็นการยากที่จะจำแนกหรือแบ่งแยกทั้ง 2 กลุ่มนี้ได้ชัดเจน

จากการทดสอบการจำแนกใบหน้าบนอุปกรณ์ฝังตัว (Raspberry Pi) สำหรับโมเดล 7 กลุ่ม และ 2 กลุ่ม โดยใช้เวลา 1 วินาที ต่อ 1 ภาพ และใช้เวลา 2 วินาที ต่อ 1 ภาพ ซึ่งสอดคล้องกับ Natalia

Efremova, Mikhail Patkin and Denis Sokolov. (2019) โดย จำแนกได้ 5 กลุ่มอารมณ์ สามารถเปรียบเทียบกับประสิทธิภาพของเครือข่ายประสาทเทียมที่ทันสมัยระบบต้องการอุปกรณ์พกพาและอุปกรณ์ที่ฝังตัว Raspberry Pi (Movidius) ได้ ซึ่งในงานวิจัยนี้มีความเร็วในการประมวลผลใช้ใกล้เคียงกัน เนื่องจากใช้เวลาในการจำแนกไม่แตกต่างกันและใช้เวลาน้อยกว่า 1 วินาที ซึ่งการรับข้อมูล (frame-rate) เข้ามาประมวลผลเพียงพอต่อการทำงานแบบ Real-time ของการจำแนกใบหน้าบนอุปกรณ์ฝังตัว

5.3 ข้อเสนอแนะ

ในการวิจัยครั้งต่อไปนำโมเดลที่ได้ไปทดลองใช้ภายในชั้นเรียน เพื่อจำแนกอารมณ์ผู้เรียน ในชั้นเรียน และนำผลมาปรับปรุงการจัดสภาพแวดล้อม ให้เกิดการเรียนรู้มากขึ้น

บรรณานุกรม

บรรณานุกรม

กันตภณ มะหาหมัด และคณะ. (2557). สภาพปัญหาและอุปสรรคต่อการจัดการเรียนการสอนฐานสมรรถนะ สำหรับการเรียนการสอนด้านเทคนิคศึกษา : กรณีศึกษาวิทยาลัยเทคนิคเพชรบุรี. *การประชุมวิชาการครุศาสตร□อุตสาหกรรมระดับชาติครั้งที่ 7*, NCTechEd07TEM08, น.260.

ฐิติรัตน์ ศิริบวรรัตนกุล (2559). *Convolutional Neural Network (CNN) เทคนิคมาตรฐานของ Deep Learning ที่นิยมใช้กันในปัจจุบัน.* สืบค้น จากhttps://mgronline.com/daily/detail/9590000091327

ธนภัทร์ คุ้มสุภา. (2559). *การจำแนกประเภทข้อความในภาษาไทยโดยใช้นิวรอลเน็ตเวิร์กคอนโวลูชันระดับตัวอักษร.* วิทยาศาสตรมหาบัณฑิต (วท.ม.) บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย.

ประเสริฐ ศักด์ อู่อรุณและคณะ (2561) *ระบบรักษาความปลอดภัยบนพื้นฐานระบบสมาร์ทโฮม: กรณีศึกษาการแจ้งเตือนเมื่อตรวจพบวัตถุเคลื่อนไหวที่เป็นมนุษย์,* Walailak Procedia 2018; 2018(5): it109

วิทูลย์ ดอนพรทัน และคณะ (2553). การหาจุดสนใจของภาพด้วย Dop-RPPRBF เพื่อหาลักษณะเด่นของภาพแบบ SIFT บนอุปกรณ์ที่มีความสามารถในการประมวลผลแบบจำกัด. *Computer Information Technologies 2010 (CIT2010),* น.130-135.

อภิวัฒน์ สวัสดิรัตน์. (2560). ระบบตรวจจับใบหน้าเพื่อลงเวลาด้วยราสเบอร์รีไพ. *การประชุมวิชาการ งานวิจัยและพัฒนาเชิงประยุกต์ ครั้งที่ 9.* สมาคมวิชาการไฟฟ้า อิเล็กทรอนิกส์ คอมพิวเตอร์ โทรคมนาคม และสารสนเทศ.

Abhishek Chaurasia, Sangpil Kim and Eugenio Culurciello, (2016). ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation. *arXiv:1606.02147v1* [cs.CV].

Andrinandrasana David Rasamoelina, Fouzia Adjailia and Peter SINC˘ AK´. (2019). Deep Convolutional Neural Network For Robust Facial Emotion Recognition. *IEEE International Symposium on INnovations in Intelligent SysTems and Applications (INISTA).* INSPEC Accession Number: 18869170, DOI: 10.1109/INISTA.2019.8778282

Alfredo Canziani, Adam Paszke, Eugenio Culurciello. (2016). An Analysis of Deep Neural Network Models for Practical Applications. *CoRR abs/1605.07678.*

Baron, R. A. (1990). Understanding and Managing the Human Side of Work Behavior in Organization. *(3 rt ed). Boston: Allyn and Bacon.*

Bharath Raj. (2018). *A Simple Guide to the Versions of the Inception Network.* Retrieved [insert date] from https://towardsdatascience.com/a-simple-guide-to-the-versions- of-the-inception-network-7fc52b863202

Bo-Kyeong Kim, Suh-Yeon Dong, Jihyeon Roh, Geonmin Kim and Soo-Young Lee (2016). Fusing Aligned and Non-Aligned Face Information for Automatic Affect Recognition in the Wild: A Deep Learning Approach**.** *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* CVPRW.

Cha Zhang and Zhengyou Zhang (2014). Improving Multiview Face Detection with Multi-Task Deep Convolutional Neural Networks*. IEEE Winter Conference on Applications of Computer Vision (WACV).* DOI: 10.1109/WACV.2014.6835990.

Challenges in Representation Learning ICML. (2013)Challenges in Represen- tation Learning: Facial Expression Recognition Challenge. *FER-2013, 2013.*

D. Viet Sang, N. Van Dat, and D. Phan Thuan, (2017). Facial expression recognition using deep convolutional neural networks. *IEEE,* DOI: 10.1109/KSE.2017.8119447.

Eugenio Culurciello. (2017). *Neural Network Architectures.* Retrieved [insert date] from https://towardsdatascience.com/neural-network-architectures- 156e5bad51ba.

Haoxiang Li and other (2015). A Convolutional Neural Network Cascade for Face Detection. *IEEE Xplore.* CVPR2015 Computer Vision.

Jaak Panksepp (1998). *Affective Neuroscience: The Foundations of Human and Animal Emotions.* New York: Oxford University Press.

Liang Zheng, Yi Yang, and Qi Tian (2015). SIFT Meets CNN: A Decade Survey of Instance Retrieval. ***J****OURNAL OF LATEX CLASS FILES, VOL. 14,* NO. 8.

Lowe, David G. (1999). Object Recognition from Local Scale-Invariant Features. *Proceedings of ICCV,* 2:1150-1157.

M. Harinthapatip. (2017) *Emotions induced by intracerebral electrical stimula- tion of the temporal lobe.* Retrieved [insert date] from https://www.ncbi.nlm.nih.gov /pubmed /17239106, 2017.

Mulder, P. (2018). *Emotion Wheel by Robert Plutchik.* Retrieved [insert date] from https://www.toolshero.com/psychology/personal-happiness/emotion-wheel- robert-plutchik/

Natalia Efremova, Mikhail Patkin and Denis Sokolov. (2019**).** Face and Emotion Recognition with Neural Networks on Mobile Devices: Practical Implementation on Different Platforms**.** *Automatic Face & Gesture Recognition (FG 2019), INSPEC Accession Number: 18821730,* DOI: 10.1109/FG.2019.8756562

Nelson, J.L., Carlson K. and Palonsky, S.B. (1993). *Critial Issues in Education : A Dialectic Approach.* New York : McGraw-Hill.

Natalia Efremova, Mikhail Patkin and Denis Sokolov. (2019). Face and Emotion Recognition with Neural Networks on Mobile Devices: Practical Implementation on Difefernt Platforms. *Automatic Face & Gesture Recognition (FG 2019), INSPEC Accession Number:* 18821730, DOI: 10.1109/FG.2019.8756562

Octavio Arriaga and Paul G. Ploger (2017). *Real-time Convolutional Neural Networks for Emotion and Gender Classification.* Hochschule Bonn-Rhcin-Sieg Department of Computer Science Grantham-Allee 20, 53757 Sankt Augustin, Germany.

P. Ekman, (1934). *Facial Action Coding System (FACS).* Retrieved [insert date] from https://www.paulekman.com/product-category /facs/.

Priya Dwivedi. (2019). *Understanding and Coding a ResNet in Keras.* Retrieved [insert date] from https://towardsdatascience.com/understanding-and-coding-a-resnet- in-keras-446d7ff84d33

W. Wan, C. Yang, and Y. Li. (2016). Facial Expression Recognition Using Convlutional Neural Network - A Case Study of the Relationship between Dataset Characteristics and Network Performance. *cs231n.*

S. Alizadeh and A. Fazel. (2017).Convolutional Neural Networks for Facial Expression Recognition. *arXiv:1704.06756, 2017.*

Syed Danish Ali and Rahul Ahuja (2016). *The Evolution and Core Concepts of Deep Learning & Neural Networks.* Retrieved [insert date] from https://www.analyticsvidhya.com/blog/2016/08/evolution-core-concepts-deep-learning-neural-networks/.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. (1998). Gradient-based learning applied to document recognition, *Proc. IEEE 86(11):* 2278–2324.

ภาคผนวก

ภาคผนวก ก

ขั้นตอนการติดตั้ง OpenCV และ TensorFlow สำหรับ Raspberry Pi

1 ติดตั้งระบบของ Raspberry Pi

Raspberry Pi จำเป็นต้องได้รับการ update โดยเปิด terminal จากนั้น Run คำสั่งต่อไปนี้

```
sudo apt-get update
sudo apt-get upgrade
```

รอสักครู่ระบบจะทำการ update & *upgrade* อาจจะใช้เวลา 10–20 นาที

หากต้องการเพิ่มพื้นที่ว่างให้กับหน่อยความจำบน Raspberry pi สามารถทำได้ โดยลบแพ็กเกจ LibreOffice และ Wolfram engine มีคำสั่งดังต่อไปนี้

```
sudo apt-get purge wolfram-engine
sudo apt-get purge libreoffice*
sudo apt-get clean
sudo apt-get autoremove
```

2 วิธีการติดตั้ง OpenCV

เริ่มต้นด้วยการ “sudo apt-get update” อีกครั้ง จากนั้นใช้คำสั่งต่อไปนี้ sudo เพื่อติดตั้ง lib ที่ Opencv มีความจำเป็นต้องใช้งาน ดังนี้

```
sudo apt-get install libjpeg-dev libtiff5-dev libjasper-dev libpng12-dev
sudo apt-get install libavcodec-dev libavformat-dev libswscale-dev libv4l-dev
sudo apt-get install libxvidcore-dev libx264-dev
sudo apt-get install libhdf5-serial-dev
sudo apt-get install qt4-dev-tools
sudo apt-get install -y libqtgui4
```

ติดตั้ง opencv-python และ opencv-contrib-python โดยใช้ pip3

```
sudo pip3 install opencv-python
sudo pip3 install opencv-contrib-python
```

ติดตั้ง driver เพื่อให้ Raspberry Pi B+ สามารถใช้งาน USB Webcam ได้โดยใช้คำสั่งนี้ (จะต้องเช็คก่อนเลือกซื้อ USB Webcam ที่ Raspberry Pi รองรับนะครับว่า https://elinux.org/RPi_USB_Webcams)

```
sudo apt-get install fswebcam
```

3 วิธีการติดตั้ง TensorFlow และแพ็คเกจพื้นฐาน

การติดตั้ง TensorFlow สำหรับ Raspberry pi นั้นจะใช้คำสั่ง “pip3” ใน การติดตั้งไลบรารีที่จำเป็นสำหรับการใช้งาน Tensorflow มีคำสั่งดังนี้

```
pip3 install Twisted
pip3 install scrapy
pip3 install pillow
pip3 install lxml
pip3 install cython
pip3 install numpy
pip3 install matplotlib
pip3 install grpcio
pip3 install h5py
```

ถัดมาจะเป็นขั้นตอนการติดตั้ง Tensorflow โดยสามารถตรวจสอบมีเวอร์ชันที่เหมาะสมกับ Raspberry Pi ดังนี้ (ลิ้งค์สำหรับเลือกเวอร์ชั่น Tensorflow ที่ต้องการ https://www.piwheels.org/simple/tensorflow/)

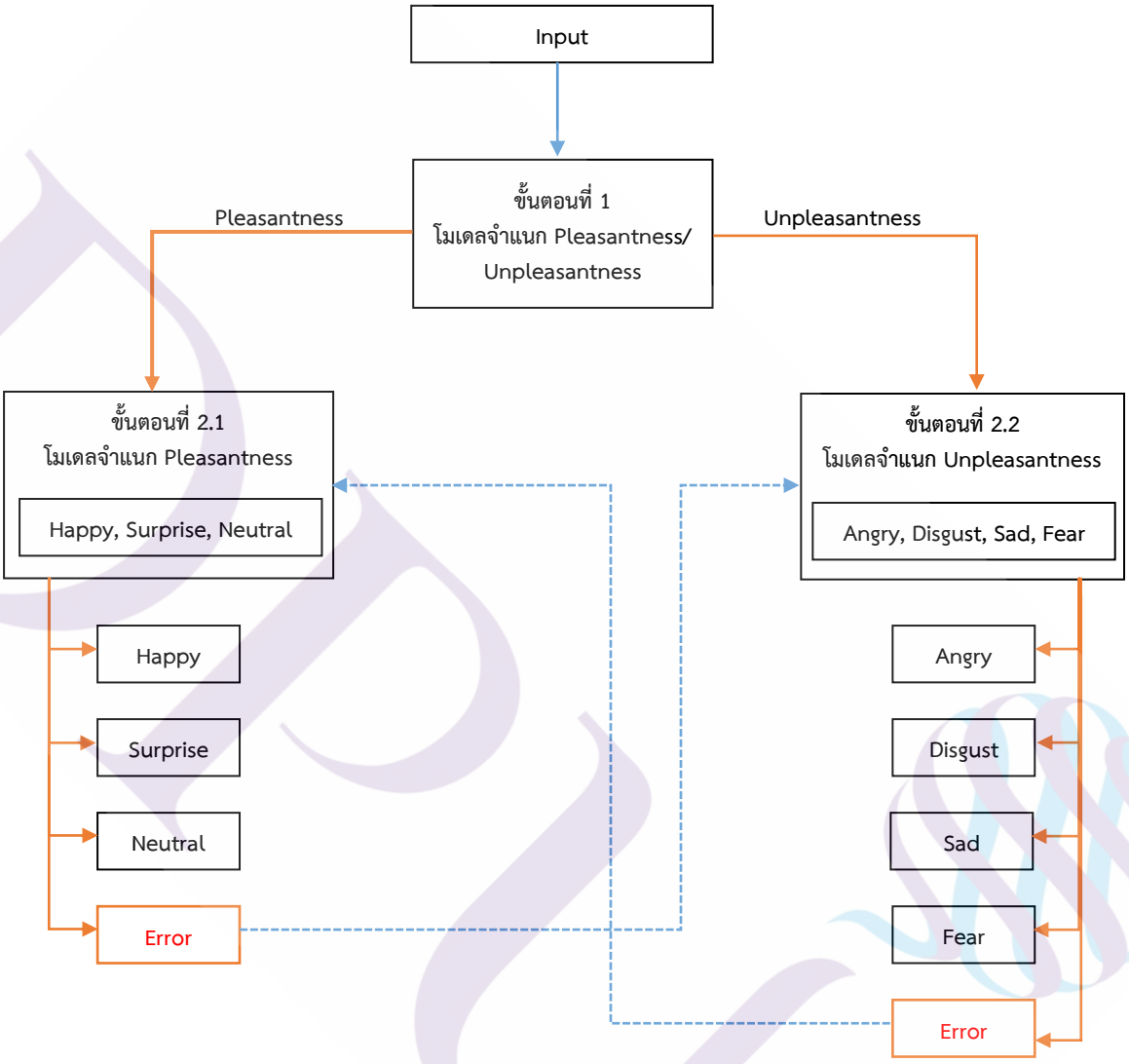
```
#คำสั่งแบบติดตั้งเวอร์ชันล่าสุด
pip3 install tensorflow
#คำสั่งแบบเลือกเวอร์ชันในการติดตั้ง
pip3 install tensorflow==1.13.1
#คำสั่งแบบโหลดไฟล์เวอร์ชั่นที่ต้องการมาติดตั้ง
wget https://www.piwheels.org/simple/tensorflow/tensorflow-1.13.1-cp37-none-linux_armv7l.whl
pip3 install /home/pi/tensorflow-1.13.1-cp37-none-linux_armv7l.whl
```

เมื่อติดตั้งเสร็จแล้วให้ทำการทดสอบโดยการเรียกใช้งาน Python3 แล้วทำการเรียกใช้งานแพคเกจ Tensorflowตามคำสั่งต่อไปนี้ :

```
python3
import tensorflow as tf
tf.__version__
```

การจัดการข้อมูลและทดสอบการจำแนก 2 steps Classification

ในขั้นตอนนี้เป็นการเตรียมข้อมูล FER 2013 เพื่อใช้ในการสร้างโมเดลแบบ 2 ขั้นตอน



ตัวอย่างโมเดลจำแนกใบหน้าบนอุปกรณ์ฝังตัว (Raspberry Pi) สำหรับ 7 กลุ่มอารมณ์ Our Mode

Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	(None, 64, 64, 1)	0	
conv2d_1 (Conv2D)	(None, 62, 62, 16)	144	input_1[0][0]
batch_normalization_1 (BatchNor	(None, 62, 62, 16)	64	conv2d_1[0][0]
activation_1 (Activation)	(None, 62, 62, 16)	0	batch_normalization_1[0][0]
conv2d_2 (Conv2D)	(None, 60, 60, 16)	2304	activation_1[0][0]
batch_normalization_2(BatchNor	(None, 60, 60, 16)	64	conv2d_2[0][0]
activation_2 (Activation)	(None, 60, 60, 16)	0	batch_normalization_2[0][0]
separable_conv2d_1 (SeparableCo	(None, 60, 60, 32)	656	activation_2[0][0]
batch_normalization_4 (BatchNor	(None, 60, 60, 32)	128	separable_conv2d_1[0][0]
activation_3 (Activation)	(None, 60, 60, 32)	0	batch_normalization_4[0][0]
separable_conv2d_2 (SeparableCo	(None, 60, 60, 32)	1312	activation_3[0][0]
batch_normalization_5 (BatchNor	(None, 60, 60, 32)	128	separable_conv2d_2[0][0]
activation_4 (Activation)	(None, 60, 60, 32)	0	batch_normalization_5[0][0]
separable_conv2d_3 (SeparableCo	(None, 60, 60, 32)	1312	activation_4[0][0]
batch_normalization_6 (BatchNor	(None, 60, 60, 32)	128	separable_conv2d_3[0][0]
activation_5 (Activation)	(None, 60, 60, 32)	0	batch_normalization_6[0][0]
separable_conv2d_4 (SeparableCo	(None, 60, 60, 32)	1312	activation_5[0][0]
batch_normalization_7 (BatchNor	(None, 60, 60, 32)	128	separable_conv2d_4[0][0]
activation_6 (Activation)	(None, 60, 60, 32)	0	batch_normalization_7[0][0]
conv2d_3 (Conv2D)	(None, 30, 30, 32)	512	activation_2[0][0]
max_pooling2d_1 (MaxPooling2D)	(None, 30, 30, 32)	0	activation_6[0][0]
batch_normalization_3 (BatchNor	(None, 30, 30, 32)	128	conv2d_3[0][0]
add_1 (Add)	(None, 30, 30, 32)	0	max_pooling2d_1[0][0] batch_normalization_3[0][0]
separable_conv2d_5 (SeparableCo	(None, 30, 30, 64)	2336	add_1[0][0]
batch_normalization_9 (BatchNor	(None, 30, 30, 64)	256	separable_conv2d_5[0][0]
activation_7 (Activation)	(None, 30, 30, 64)	0	batch_normalization_9[0][0]
separable_conv2d_6 (SeparableCo	(None, 30, 30, 64)	4672	activation_7[0][0]
batch_normalization_10 (BatchNo	(None, 30, 30, 64)	256	separable_conv2d_6[0][0]
activation_8 (Activation)	(None, 30, 30, 64)	0	batch_normalization_10[0][0]
conv2d_4 (Conv2D)	(None, 15, 15, 64)	2048	add_1[0][0]
max_pooling2d_2 (MaxPooling2D)	(None, 15, 15, 64)	0	activation_8[0][0]
batch_normalization_8 (BatchNor	(None, 15, 15, 64)	256	conv2d_4[0][0]
add_2 (Add)	(None, 15, 15, 64)	0	max_pooling2d_2[0][0] batch_normalization_8[0][0]
separable_conv2d_7 (SeparableCo	(None, 15, 15, 128)	8768	add_2[0][0]
batch_normalization_12 (BatchNo	(None, 15, 15, 128)	512	separable_conv2d_7[0][0]

activation_9 (Activation)	(None, 15, 15, 128)	0	batch_normalization_12[0][0]
separable_conv2d_8 (SeparableCo	(None, 15, 15, 128)	17536	activation_9[0][0]
batch_normalization_13 (BatchNo	(None, 15, 15, 128)	512	separable_conv2d_8[0][0]
activation_10 (Activation)	(None, 15, 15, 128)	0	batch_normalization_13[0][0]
conv2d_5 (Conv2D)	(None, 8, 8, 128)	8192	add_2[0][0]
max_pooling2d_3 (MaxPooling2D)	(None, 8, 8, 128)	0	activation_10[0][0]
batch_normalization_11 (BatchNo	(None, 8, 8, 128)	512	conv2d_5[0][0]
add_3 (Add)	(None, 8, 8, 128)	0	max_pooling2d_3[0][0]
			batch_normalization_11[0][0]
separable_conv2d_9 (SeparableCo	(None, 8, 8, 256)	33920	add_3[0][0]
batch_normalization_15 (BatchNo	(None, 8, 8, 256)	1024	separable_conv2d_9[0][0]
activation_11 (Activation)	(None, 8, 8, 256)	0	batch_normalization_15[0][0]
separable_conv2d_10 (SeparableC	(None, 8, 8, 256)	67840	activation_11[0][0]
batch_normalization_16 (BatchNo	(None, 8, 8, 256)	1024	separable_conv2d_10[0][0]
activation_12 (Activation)	(None, 8, 8, 256)	0	batch_normalization_16[0][0]
conv2d_6 (Conv2D)	(None, 4, 4, 256)	32768	add_3[0][0]
max_pooling2d_4 (MaxPooling2D)	(None, 4, 4, 256)	0	activation_12[0][0]
batch_normalization_14 (BatchNo	(None, 4, 4, 256)	1024	conv2d_6[0][0]
add_4 (Add)	(None, 4, 4, 256)	0	max_pooling2d_4[0][0]
			batch_normalization_14[0][0]
separable_conv2d_11 (SeparableC	(None, 4, 4, 512)	133376	add_4[0][0]
batch_normalization_18 (BatchNo	(None, 4, 4, 512)	2048	separable_conv2d_11[0][0]
activation_13 (Activation)	(None, 4, 4, 512)	0	batch_normalization_18[0][0]
separable_conv2d_12 (SeparableC	(None, 4, 4, 512)	266752	activation_13[0][0]
batch_normalization_19 (BatchNo	(None, 4, 4, 512)	2048	separable_conv2d_12[0][0]
activation_14 (Activation)	(None, 4, 4, 512)	0	batch_normalization_19[0][0]
conv2d_7 (Conv2D)	(None, 2, 2, 512)	131072	add_4[0][0]
max_pooling2d_5 (MaxPooling2D)	(None, 2, 2, 512)	0	activation_14[0][0]
batch_normalization_17 (BatchNo	(None, 2, 2, 512)	2048	conv2d_7[0][0]
add_5 (Add)	(None, 2, 2, 512)	0	max_pooling2d_5[0][0]
			batch_normalization_17[0][0]
separable_conv2d_13 (SeparableC	(None, 2, 2, 786)	407040	add_5[0][0]
batch_normalization_21 (BatchNo	(None, 2, 2, 786)	3144	separable_conv2d_13[0][0]
activation_15 (Activation)	(None, 2, 2, 786)	0	batch_normalization_21[0][0]
separable_conv2d_14 (SeparableC	(None, 2, 2, 786)	624870	activation_15[0][0]
batch_normalization_22 (BatchNo	(None, 2, 2, 786)	3144	separable_conv2d_14[0][0]
activation_16 (Activation)	(None, 2, 2, 786)	0	batch_normalization_22[0][0]
conv2d_8 (Conv2D)	(None, 1, 1, 786)	402432	add_5[0][0]
max_pooling2d_6 (MaxPooling2D)	(None, 1, 1, 786)	0	activation_16[0][0]
batch_normalization_20 (BatchNo	(None, 1, 1, 786)	3144	conv2d_8[0][0]

add_6 (Add)	(None, 1, 1, 786)	0	max_pooling2d_6[0][0]	batch_normalization_20[0][0]
conv2d_9 (Conv2D)	(None, 1, 1, 7)	49525	add_6[0][0]	
global_average_pooling2d_1 (Glo	(None, 7)	0	conv2d_9[0][0]	
predictions (Activation)	(None, 7)	0	global_average_pooling2d_1[0][0]	

==========

Total params: 2,222,419

Trainable params: 2,211,559

Non-trainable params: 10,860

total flops (model_1): 115,186,970

ภาคผนวก ข

ผลงานวิจัย

Facial Expression Classification using Deep Extreme Inception Networks

Thitiphong Raksarikorn and Thanapat Kangkachit
College of Innovative Technology and Engineering
Dhurakij Pundit University
Bangkok, Thailand
595162020009@dpu.ac.th, thanapat.kan@dpu.ac.th

***Abstract*—Facial expression classification plays crucial role in human-computer interaction. A large number of automated methods have been proposed since the past decades. Recently, deep learning is broadly applied in computer vision field as well as facial expression classification. The reasons are to avoid complex feature extraction process and obtained satisfied classification performance. In this work, we propose a deep convolutional neural networks (CNNs) model, inspired from XCEPTION, to classify seven groups of facial expressions. To efficiently use of model parameters, the model architecture has only 2.2 million parameters which is about 10 times less than XCEPTION. The experimental results on FER-2013 dataset show that our model offers comparable accuracy (0.7169) to the state-of-the-art methods and the upper-bound level of human accuracy (0.65±5). In addition, our model uses less number of parameters than the state-of-the-art models and without using extra features and data augmentation.**

***Index Terms*—emotions, facial expression classification, computer vision, convolutional neural networks, XCEPTION**

I. Introduction

Facial expression is a vital element of human communication. Indeed, facial expression as a form of nonverbal communication can be interpreted substantially faster than verbal communication. The human emotions can also be investigated over facial expressions during regular interaction between human beings. The well-known and broadly used of facial expression definition was defined in [1] called Facial Action Coding System (FACS). FACS defined seven groups of facial expressions which are happiness, sad, anger, fear, surprise, disgust and contempt. The automated methods to recognize these facial expressions provide a new dimension to human-computer interactions (HCI). The examples of such applications are customer satisfaction during receiving service; student intention in the classroom; the appropriate reactions by robot [2]. For such applications, facial expression classification becomes an crucial problem in computer vision research field.

Recently, convolutional neural networks (CNNs) [3] have been broadly applied to classify images instead of using the model induction based approach. To come across the complex feature extraction, CNNs uses several convolutional layers to learn features from those input images. The coarse features are learned during the first couple layers of the network. Meanwhile, the more complex features are captured in the deeper layers of the network. Generally, CNNs-based approaches provide satisfied classification performance.

978-1-5386-5538-2/18/$31.00 ©2018 IEEE

Regarding to facial expression recognition in the wild challenge holded by Kaggle in 2013, [4] provided the best accuracy about 71.9% on FER-2013 dataset [5]. This method was based on CNNs trained with square hinged loss. The tranfer learning approach was applied to CNN architectures in [6]. This work performed fine-tuning on a pre-trained generic network (i.e. for ImageNet dataset) using facial expressions dataset. This work also provided better accuracy than the challenge baseline. Later, [7] proposed a hybrid method to merge Dense SIFT features (Scale Invariant Feature Transform) with CNNs features. To the best of our knowledge, this method generated state-of-art accuracy on FER-2013 dataset which is 73.4%. Recently, [8] proposed a CNNs model named BK-VGG12 inspired from VGG architecture. This model produced accuracy of 71.9%. Interestingly, the number of parameters was decreased to 4.19 million parameters due to the use of 2048 nodes in the two fully connected (FC) layers. However, VGG-like model may not efficiently use of model parameters as in XCEPTION-like model.

In this work, we propose a deep CNNs model, inspired from XCEPTION, to classify seven groups of facial expressions. The proposed model utilizes both residual modules and depthwise separable convolutions to decouple the mapping of cross-channels and spatial correlations. More specifically, our network architecture is less complex compared to XCEPTION. Our model aims to learn more global or coarse features since the small size of input images (64x64 pixels) allows us to do more calculation with small extra cost. The global features may help to increase classification performance in cases of ambiguity between classes such as sad, fear and neutral. As result, our obtained recall for FER-2013 dataset increases dramatically while retains satisfied precision.

The rest of paper is organized as follows. In section II, several related works about facial expressions classification are described. Section III provides more detail about our network architecture. In Section IV, experimental results are illustrated and discussed. Finally, we conclude in Section V.

II. Related Work

Although, there are a large number of researches related to facial expression classification since the past decades. In

this section, we focus on several researches that applied deep learning to mitigate this problem on FER-2013 dataset that was used in the competition held by Kaggle in 2013.

CNNs is arguably the most popular deep learning architectures. In the past few years, several effective CNNs architectures have been proposed i.e. VGG-19 [9], InceptionV3 [10] and XCEPTION [11]. In general, most of the parameters in CNNs are in fully connected layer which uses for classification aspect as in VGG-19. InceptionV3 reduces the amount of parameters by using global average pooling which takes the average over all elements in each feature map. Meanwhile, XCEPTION utilizes the advantages of using residual modules [10] and depthwise separable convolutions. As result, XCEPTION produced the best performance on the ImageNet dataset and also on a large image dataset comprising of 350 million images and 17,000 classes. Although, XCEPTION has nearly the same number of parameters as Inception V3 (i.e. 22.85 million and 23.65 million respectively). XCEPTION has claimed to efficiently use the model parameters than Inception V3. However, XCEPTION model may be more complicated for facial expression classification which contains only seven groups of expressions and small size images (64x64 pixels) as in FER-2013 dataset [5].

The winner of FER-2013 facial expression recognition competition is RBM team [12] with the accuracy of 69.4% on public test set and 71.2% on private test set. To the best of our knowledge, this is the state-of-the-art on the FERC-2013 dataset so far. This work uses CNNS that includes 3 convolutional layers (with large filter size). There is only one maxpool layer after the first convolutional layer. For classification aspect, there are 2 fully connected layers, where the first has 3072 nodes and the second is the output layer with 7 nodes. Instead of softmax activation function, the multi-class SVM loss is applied to output layer.

Recently, according to the effective of VGG architecture in image classification problem, [8] proposed VGG-like architecture to classify facial expressions. The BKVGG-12 and BKVGG-14 models are presented with 12 and 14 convolutional layers respectively. There are also 3 fully connected layers, where the first and second have 256 nodes and the third has 7 nodes. In addition, this work uses data augmentation and L2 multi-class SVM loss function. This work claims to produce 71.0% accuracy on public test set and 71.9% accuracy on private test set by using BKVGG-12. However, the BKVGG-12 and VGG-14 architectures have about 4.19 M and 4.92 M parameters respectively. However, the number of parameters is still larger than the number of parameters of our model.

A hybrid method to classify facial expressions was proposed in [7]. This work provided the state-of-art results i.e.73.4% accuracy on FER-2013 and 99.1% accuracy on CK+ dataset. Scale Invariant Feature Transform (SIFT) features are used to increase performance on small data. Both regular SIFT and Dense SIFT were merged with CNNs that produced the state-of-art accuracy.

III. Classification of Facial Expressions

Our proposed model is based on convolutional neural networks (CNNs) architectures. More specifically, this model entirely utilizes depthwise separable convolution layers as inspired from XCEPTION model. Thus, the hypothesis of this model is that the mapping of cross-channels correlations and spatial correlations can be entirely decoupled.

A complete architecture of the network is depicted in Fig. 1. The network is composed of 3 flows. The input data first enters the entry flow, and then through the middle flow which is repeated four times with increasingly number of filters each time and finally through the exit flow. Hence, all separable convolutional layers also have no depth expansion as the same as in XCEPTION model. This architecture is composed of 14 convolutional layers as feature extraction for the network.

In the **entry flow**, coarse features are extracted through 4 convolutional layers followed by batch normalization and RELU activation function. Note that the number of filters always increases two times from the former convolutional layers.

In the **middle flow**, more complex features are extracted through 8 separable convolutional layers. After passing through 2 separable convolutional layers, the number of filters increases two times as well i.e. starting from 64 filters, then 128 filters, then 256 filters and finally 512 filters. Hence, at the end of the middle flow, we will have 512 features of 2x2 mapping.

In the **exit flow**, the most detailed features are extracted through 2 separable convolutional layers with 786 filters. Then the last convolutional layer is applied with 7 filters of 3x3 regarding to 7 groups of facial expression. Then, global average pooling is use to reduce the mapping size of 3x3 to 1x1. Finally, softmax activation function is applied to classify facial expressions.

IV. Experimental Results

A. Setup

The experiments are conducted on FER-2013 dataset which is a popular standard dataset for facial expression classification. The dataset consists of 35,887 gray images of 48x48 resolution. The training dataset consists of 28,709 images while the rests are validated and test data which are 3,589 images for each dataset. Each image contains a human face (in the wild) with one from seven facial expressions i.e. angry, disgust, fear, happy, sad, surprise and neutral. The examples of FER-2013 dataset are shown in Fig. 2.

The classification performance is measured through several measurements (i.e. accuracy, precision and recall) on the test dataset of FER-2013. We also compare classification performance of our proposed model with several CNNs-based models i.e. mini XCEPTION [13], BKVGG-12, BKVGG-14 [8] and XCEPTION [10]. In addition, we also compare with several CNNs-based models introduced in the source codes of [13] which are simple CNN, simpler CNN, tiny XCEP-TION and big XCEPTION.

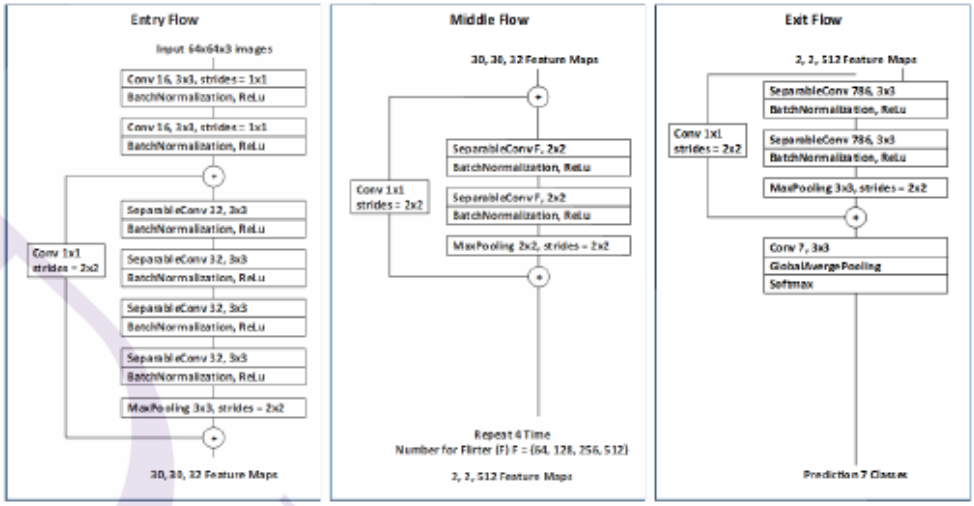


Figure 1: Our model architecture inspired from XCEPTION model.

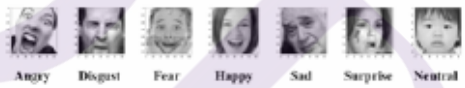


Figure 2: Samples of FER-2013 dataset.

This work is implemented using Keras [14] running on top of Tensor-Flow with GPU support (NVIDIA GTX 1070). The model is trained using training dataset for 500 epochs using ADAM optimizer [15]. After finish training, only the most accurate model evaluated on validation set is chosen.

B. Experimental Results of Our Model

The precision and recall for each group of facial expressions produced by our CNNs model is illustrated in Fig. 3. The overall precision and recall are also plotted as dotted lines in the graph. Moreover, confusion matrix is presented in Fig. 4.

The overall precision and recall produced by our CNNs model are 0.7291 and 0.7077 as shown in Fig. 3. There are 3 groups of facial expressions i.e. happy, surprise, disgust that produce higher precision and recall than the overalls. As expected, the highest precision and recall are obtained by classifying the happy face due to the large number of training faces. When classifying surprise and disgust faces, obtained precision and recall are also high since both faces have quite clear visual characteristics and does not hard to distinguish them from the others. Thus, CNNs can learn the distinctive characteristics that lead to accurate classification.

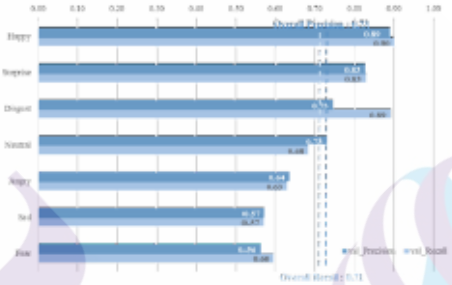
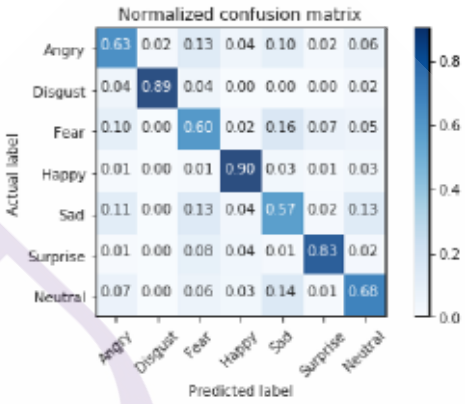


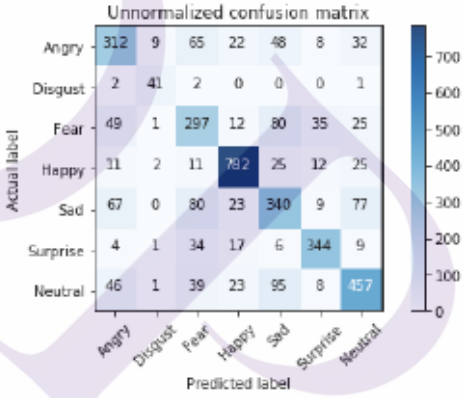
Figure 3: Precision and recall obtained by our model.

In case of classifying neutral and angry faces, obtained recall of both faces is almost equal but lower than the overall precision and recall. However, precision of classifying neural faces is significantly better than that of angry faces. It may due to number of training faces. As shown in Fig. 4(a) and Fig. 4(b), it seems that a large portion (about 14%) of neutral faces are misclassified as sad face. In contrast, a large portion of angry faces are misclassified as fear and sad faces (i.e. 13% and 10% respectively).

In case of classifying sad and fear faces, obtained precision of both faces is comparable but also significantly lower than the overall precision. However, recall of classifying fear face is significantly better than that of sad face. We can notice that a large portion of fear faces are misclassified as sad and angry



(a)



(b)

Figure 4: (a) Normalized confusion matrix of our model (b) Unnormalized confusion matrix of our model.

faces (i.e. 16% and 10% respectively) as shown in Fig. 4(b). In addition, the samples of correctly and incorrectly classified images of sad and fear faces are shown in Fig. 5. Noticed that even human beings are not capable of distinguish these faces.

C. Comparison Results With Other Models

As clearly seen in Table I, our model offers significantly higher accuracy and recall but comparable in precision compared to the other models. The reason is that our model is more complex (i.e. having more parameters) than the others. Thus, our model can learn more complex distinctive features for each facial expression than the other models.

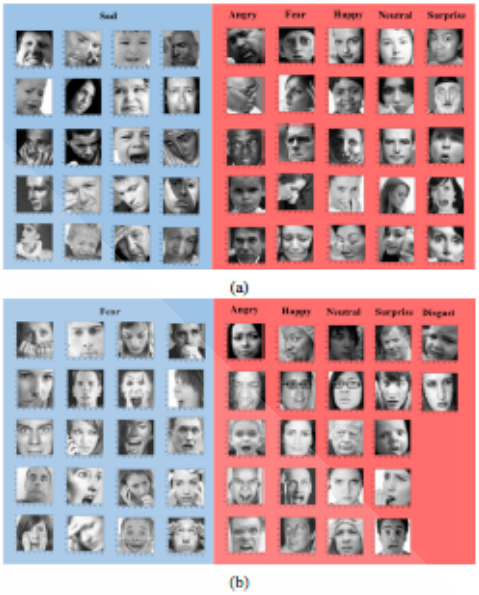


Figure 5: Samples of correctly (left) and incorrectly (right) classified images (a) classified as "Sad" (b) classified as "Fear".

Table I: Performance Comparison With State-of-the-art Models

Model	Accuracy	Precision	Recall	#Params*
Simple CNN	0.6291	0.7524	0.4954	0.64
Simpler CNN	0.6356	0.7627	0.4976	0.60
Tiny XCEPTION	0.6236	0.7392	0.5056	0.02
Mini XCEPTION	0.6601	0.7462	0.5644	0.06
Big XCEPTION	0.6606	0.7406	0.5836	0.21
BKVGG-12	0.7100	N/A	N/A	4.19
BKVGG-14	0.7060	N/A	N/A	4.92
XCEPTION	0.7144	0.7210	0.7054	20.87
Our Model	**0.7169**	**0.7291**	**0.7077**	**2.22**

*in million parameters

Although the human accuracy for classifying facial expression is 65±5% [4]. Our model performs in almost the upper-bound value of human-level performance. In another word, our model performance is comparable to human performance.

Compare to the winner model [4] in the challenge using FER-2013 dataset which generated 71.2% accuracy, our model achieves a bit higher accuracy of 71.69%. In addition, our model uses less parameters (i.e. 2.2 vs 5 million). In other words, our model produces comparable accuracy but less complex. When compare to the-state-of-art accuracy (73.4%) in [7] that merges SIFT features with CNNs, our model

produce a bit lower accuracy. Unlikely, our model uses the raw data from FER-2013 dataset without using extra features and data augmentation.

V. Conclusion and Future Works

In this paper, the problem of classifying facial expression is addressed. To deal with this problem, we propose a CNNs model inspired from XCEPTION model. The model contains around 2.2 million parameters without using fully connected layers. The model produces classification accuracy in almost highest-upper bound of human-level performance measured in the standard dataset i.e. FER-2013. The model also provides comparable accuracy with the-state-of-the-art models but using less parameters and without using extra features.

In the future, data preprocessing and data augmentation will be applied to FER-2013 dataset to increase number of training images. As result, the model will capable of handling more variety of input images resulted in increasing of classification accuracy.

References

[1] P. Ekman, "Facial Action Coding System (FACS)." [Online] Discovered on July 15, 2017. From https://www.paulekman.com/product-category/facs/, 1934.

[2] M. Harinthapatip, "Emotions induced by intracerebral electrical stimulation of the temporal lobe."[Online] Discovered on July 15, 2017. https://www.ncbi.nlm.nih.gov/pubmed/17239106, 2017.

[3] S. Alizadeh and A. Fazel, "Convolutional Neural Networks for Facial Expression Recognition," arXiv:1704.06756, 2017.

[4] Y. Tang. "Deep learning using linear support vector machines." ArXiv preprint arXiv: 1306.0239, 2013.

[5] Challenges in Representation Learning ICML. "Challenges in Representation Learning: Facial Expression Recognition Challenge. FER-2013 .", 2013.

[6] H.-Wei Ng, V. Dung Nguyen, V. Vonikakis and S. Winkler, "Deep Learning for Emotion Recognition on Small Datasets Using Transfer Learning", ICMI '15 Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Pages 443-449. 2015.

[7] M. Al-Shabi, W. Ping Cheah and T. Connie, "Facial Expression Recognition Using a Hybrid CNN–SIFT Aggregator", Springer International Publishing AG. Part of Springer Nature. 2017.

[8] D. Viet Sang, N. Van Dat, and D. Phan Thuan, "Facial expression recognition using deep convolutional neural networks". IEEE, DOI: 10.1109/KSE.2017.8119447, 2017.

[9] K. Simonyan, and A. Zisserman, et al: "Very Deep Convolutional Networks for Large-Scale Image Recognition" Subjects: Computer Vision and Pattern Recognition (cs.CV) Cite as:arXiv: 1409.1556, 2015.

[10] C. Szegedy, and V. Vanhoucke, "Rethinking the Inception Architecture for Computer Vision." arXiv:1512.00567, 2015.

[11] F. ois Chollet, "XCEPTION: Deep learning with depthwise separable convolutions." CoRR, abs/1610. 02357, 2016.

[12] W. Wan, C. Yang, and Y. Li. "Facial Expression Recognition Using Convlutional Neural Network - A Case Study of The Relationship Between Dataset Characteristics and Network Performance". cs231n. 2016

[13] O. Arriaga and P.G. Ploger, "Real-time Convolutional Neural Networks for Emotion and Gender Classification." Hochschule Bonn-Rhein-Sieg Department of Computer Science Grantham-Allee 20, 53757 Sankt Augustin, Germany, 2017.

[14] F. Chollet. "Keras: Deep Learning library for Theano and TensorFlow." [Online] Discovered on July 15, 2017. From https://faroit. github.io/keras-docs/1.2.0/, 2015.

[15] D. Kingma, and J. Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980, 2014.

ประวัติผู้เขียน

ชื่อ-นามสกุล	ฐิติพงษ์ รักษาริกรณ์
ประวัติการศึกษา	พ.ศ. 2559 ปริญญาตรี (ค.บ. 5 ปี) สาขาวิชาคอมพิวเตอร์ศึกษา คณะครุศาสตร์ มหาวิทยาลัยราชภัฏนครสวรรค์
ตำแหน่งและสถานที่ทำงานปัจจุบัน	ครูผู้ช่วย วิชาเอกคอมพิวเตอร์ โรงเรียนร่องตาทีวิทยา สังกัดสำนักงานเขตพื้นที่การศึกษามัธยมศึกษา เขต 42