บทที่ 2

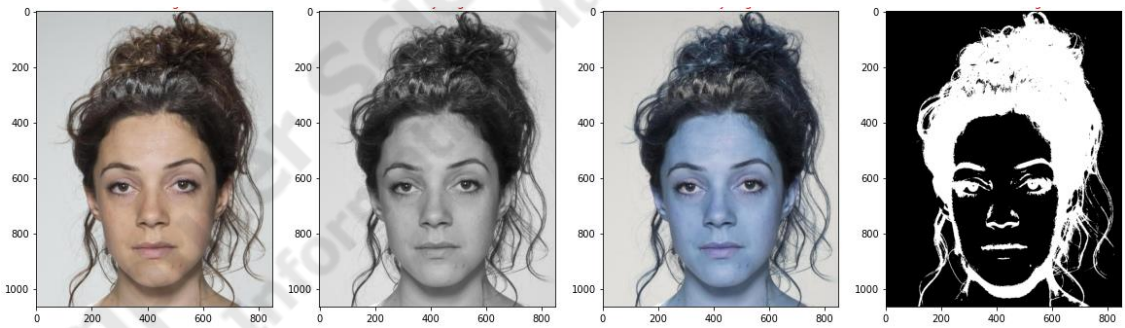
ทฤษฎีและระบบงานที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 การประมวลผลภาพ

การประมวลผลภาพ (Image Processing) [1] คือกระบวนการประมวลผลภาพที่ถูกป้อนเข้ามาในระบบ เพื่อทำการปรับปรุงคุณภาพ วิเคราะ ฟื้นฟูรายละเอียด หรือคัดเลือกรายละเอียดข้อมูลที่ต้องการจากภาพที่ถูกป้อนเข้ามา การประมวลผลภาพมีอยู่ 2 วิธี คือ ดิจิทัล (Digital) และ อนาล็อก (Analog) ในที่นี้จะกล่าวถึงการประมวลผลภาพแบบ Digital เป็นหลัก เพื่อใช้ในการจัดการข้อมูลภาพที่เข้ามาในระบบ

โดยทั่วไปแล้วการประมวลผลภาพ Digital ระบบจะมองภาพเป็น สัญญาณ 2 มิติ (two-Dimensional Signals) แล้วจึงเริ่มกระบวนการประมวลผลด้วยวิธีที่ถูกกำหนดไว้โดยผู้สร้างโปรแกรม เพื่อให้ได้ส่วนข้อมูลเฉพาะที่ต้องการจากภาพเหล่านั้น ข้อมูลที่ถูกป้อนเข้าไปในระบบจะเป็นภาพนิ่ง ไม่ว่าจะเป็นภาพที่ได้จากการถ่ายภาพ หรือภาพบางส่วนที่นำออกมาจากวิดีโอ ส่วนข้อมูลที่ได้จากการประมวลผลภาพ จะเป็นรูปแบบข้อมูลที่ถูกกำหนดไว้แล้วว่าต้องการสิ่งใดจากภาพที่ส่งเข้าไป



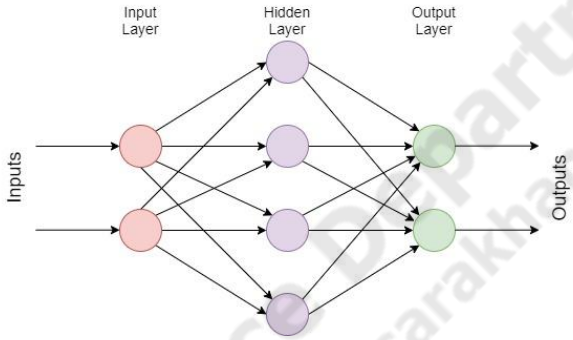
ภาพประกอบที่ 2.1 ตัวอย่างการทำ Image Processing

2.1.2 โครงข่ายประสาทเทียม

โครงข่ายประสาทเทียม (Artificial Neural Network, ANN) [2] ซึ่งจะเป็นพื้นฐานของการพัฒนาโมดูลการจำลองภาพ ซึ่ง ANN คือรากฐานของปัญญาประดิษฐ์ซึ่งเป็นชุดอัลกอริทึมที่มีวัตถุประสงค์ในการเรียนรู้ความสัมพันธ์ของชุดข้อมูลต่าง ๆ ชุดอัลกอริทึมนี้จะมีหลักการทำงานที่คล้ายคลึงคับโครงข่ายประสาทของสมองมนุษย์ กล่าวคือ มีการรวมกลุ่มของหน่วยประมวลผลย่อยหลาย ๆ หน่วย ซึ่ง ANN สามารถปรับเปลี่ยนวิธีการในรับมือกับข้อมูล Input ได้ตลอดเวลา เพื่อให้สามารถจำลองผลลัพธ์ที่ดีที่สุดได้โดยที่ไม่จำเป็นต้องเปลี่ยนแปลงกฎเกณฑ์ในการวิเคราะและแสดงผลลัพธ์ของอัลกอริทึม

หน่วยพื้นฐานของ ANN เรียกว่า นิวรอน, โหนด หรือหน่วย (Neuron, Node or Unit) ซึ่งแต่ละโหนดจะเป็นเป็นตัวที่คอยรับข้อมูลจากโหนดอื่นๆ หรือจากภายนอกโครงข่าย ข้อมูลที่รับเข้า (Input) จะถูกวิเคราะห์ และกำหนดค่าน้ำหนัก (Weight) ที่ใช้สำหรับการระบุลักษณะของรูปแบบของข้อมูลนั้น ๆ และใช้เพื่ออ้างอิงให้กับ Input ตัวอื่น ๆ ที่มีค่า Weight ที่เท่ากันหรือใกล้เคียง

ตัวอย่างชุดอัลกอริทึมที่มีรูปแบบการทำงานของ ANN เช่น โครงข่ายประสาทแบบส่งผ่าน (Feedforward Neural Network, FNN) เป็นชุด Algorithm รูปแบบแรกที่ถูกออกแบบขึ้น โดยมีการจัดเรียงโหนดแบบเป็นชั้น ๆ (Layers) และโหนดในแต่ละชั้นจะมีการเชื่อมต่อ (Connections) หรือเรียกว่าเส้นเชื่อม (Edges) กับชั้นอื่น ๆ และแต่ละเส้นเชื่อมนั้นก็จะมีค่าน้ำหนัก (Weight) เฉพาะตัว



ภาพประกอบที่ 2.2 ตัวอย่างของ Feed-Forward Neural Network

โครงข่ายประสาทแบบส่งผ่าน มีชนิดของโหนดอยู่ 3 ชนิด ดังนี้

1. โหนดรับข้อมูล (Input Nodes) เป็นส่วนที่ทำหน้าที่รับข้อมูลจากภายนอกของระบบ อยู่ในชั้นที่ทำหน้าที่รับข้อมูล (Input Layer) ในชั้นนี้จะไม่มีการประมวลผลใดๆ ทำหน้าที่เพียงรับข้อมูลและรอการส่งผ่านข้อมูลไปยังโหนดซ่อนเร้น

2. โหนดซ่อนเร้น (Hidden Nodes) โหนดซ่อนเร้นจะอยู่ในชั้นซ่อนเร้น (Hidden Layer) ซึ่งในโครงข่ายประสาทแบบส่งผ่านสามารถมีชั้นซ่อนเร้นได้มากกว่า 1 ชั้น หรือไม่มีเลย แต่จะมีชั้นรับข้อมูลและส่งออกเพียงอย่างละชั้นเท่านั้น โหนดซ่อนเร้นจะไม่มีการติดต่อกับภายนอกระบบ โหนดชนิดนี้จะทำการประมวลผลข้อมูลที่ได้รับมาจากโหนดรับข้อมูล เมื่อการทำงานของโหนดซ่อนเร้นสำเร็จ จะทำการส่งผ่านผลลัพธ์ของการประมวลผลไปยังโหนดส่งออก

3. โหนดส่งออก (Output Nodes) โหนดส่งออกจะอยู่รวมกันในชั้นส่งออก (Output Layer) ทำหน้าที่ในการประมวลผลข้อมูลที่ได้รับมาจากโหนดซ่อนเร้น หรือจากโหนดรับข้อมูลโดยตรง เมื่อการประมวลผลสำเร็จ โหนดส่งออกจะทำการส่งข้อมูลออกไปยังโครงข่าย

2.1.3 การเรียนรู้เชิงลึก

การเรียนรู้เชิงลึก (Deep Learning) [3] เป็นส่วนหนึ่งของการเรียนรู้ของอุปกรณ์ (Machine Learning) [3] ซึ่งอยู่ในกระบวนการทำงานของปัญญาประดิษฐ์ (Artificial Intelligent, AI) [3] โดยการทำงานของ Deep Learning ได้แรงบันดาลใจมาจากการทำงานของสมองมนุษย์ อัลกอริทึม Deep Learning จะมีการเรียนรู้และสะสมความรู้ (Knowledge) ด้วยตัวเองจากข้อมูลที่ถูกป้อนเข้ามา การ

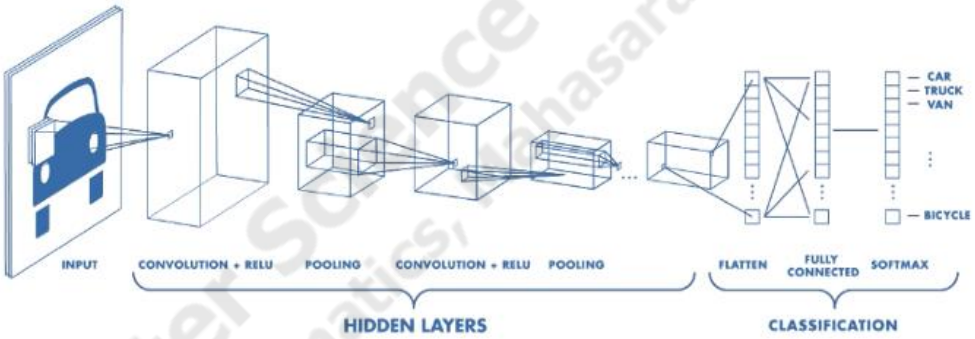
เรียนรู้ของ Deep Learning จะทำงานในรูปแบบโครงข่ายประสาท (Neural Network) มาใช้ในการวิเคราะห์ข้อมูลต่าง ๆ

2.1.4 โครงข่ายประสาทแบบคอนโวลูชันนอล

โครงข่ายประสาทแบบคอนโวลูชันนอล (Convolutional Neural Network, CNN) [4] [5] จะถูกนำมาใช้เป็น Framework อ้างอิงในการพัฒนา Generator Model

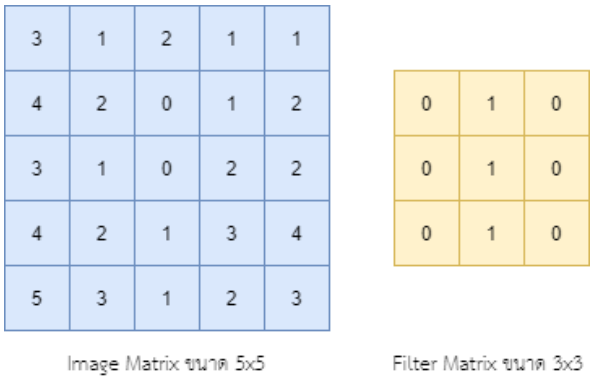
CNN เป็นรูปแบบหนึ่งของโครงข่ายประสาทเทียมที่ถูกใช้งานอย่างกว้างขวางในด้านการประมวลผลภาพและการระบุวัตถุที่ต้องการจากภาพ CNN หลักการของ CNN คือการนำภาพไปผ่านตัวกรอง (Filter) ในชั้นต่างๆ แล้วสุดท้ายจะนำข้อมูลที่ได้จากการผ่านตัวกรองในชั้นต่าง ๆ ไปเข้าฟังก์ชัน SoftMax เพื่อทำการระบุความเป็นไปได้ของวัตถุภายในภาพ

CNN จะทำการประมวลผลภาพในวิธีที่ถูกกำหนดไว้ในแต่ละชั้น ชั้นการประมวลผลมีอยู่ 3 ชั้น คือ Convolutional Layer, Pooling Layer และ Fully Connected layer ซึ่งแต่ละชั้นจะมีการประมวลผลที่แตกต่างกันในการคำนวณและค้นหาคุณลักษณะของภาพที่เหมาะสมเพื่อใช้ในการคัดแยกรูปแบบของภาพ สามารอธิบายวิธีการทำงานของชั้นต่าง ๆ ได้ดังนี้

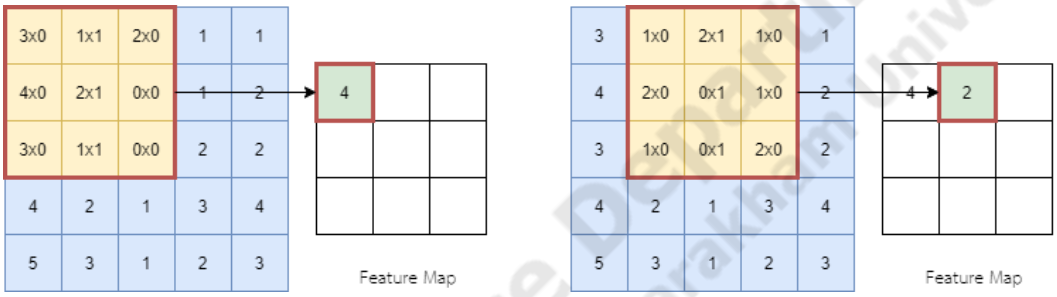


ภาพประกอบที่ 2.3 โครงสร้างของโครงข่ายประสาทแบบ CNN
ที่มา: https://towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939

1. Convolutional Layer เป็นชั้นแรกของ CNN ทำหน้าที่ในการคัดแยกคุณลักษณะของภาพที่ถูกนำเข้ามา โดยการนำข้อมูลของภาพมาคำนวนกับ แมทริกซ์ตัวกรอง (Filter Matrix) ที่ถูกกำหนดไว้ ข้อมูลที่ถูกคำนวนได้ในแต่ละครั้งจะถูกนำมาเก็บไว้ใน ผังฟีเจอร์ (Feature Map) จนกระทั่งผังฟีเจอร์เต็ม ยกตัวอย่างภาพขนาด 6x6 pixel และ Filter Matrix ขนาด 3x3 pixel เมื่อนำมาคำนวนแล้วจะได้ Feature Map ขนาด 6x6 โดยการคำนวณ Feature Map จะเป็นการนำ Filter Matrix มาวางทาบ แมทริกซ์ภาพ (Image Matrix) ที่มุมใดมุมหนึ่ง ทำการคำนวณหาผลลัพธ์ของจุด ๆ นั้น แล้วจึงเก็บผลลัพธ์ไว้ใน Feature Map ตัวกรองจะเลื่อนพิกเซลไปเรื่อย ๆ จนกว่า Feature Map จะถูกเติมครบทุกช่อง

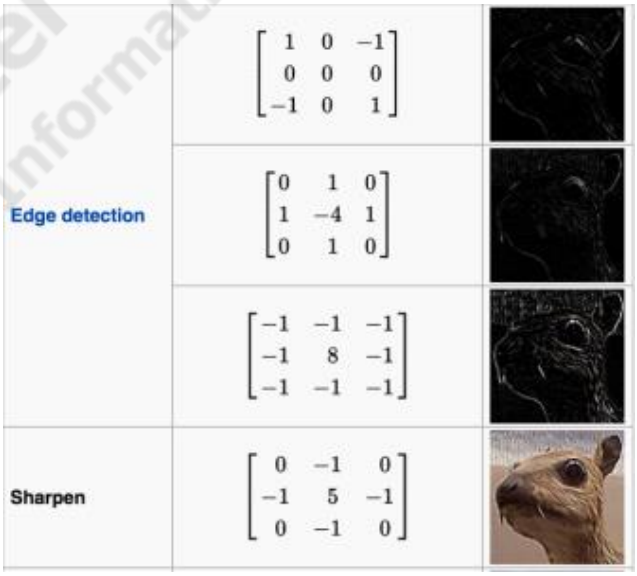


ภาพประกอบที่ 2.4 ตัวอย่าง Image Matrix (ซ้าย) และ Filter Matrix (ขวา)



ภาพประกอบที่ 2.5 ตัวอย่างการคำนวณด้วย Filter Matrix

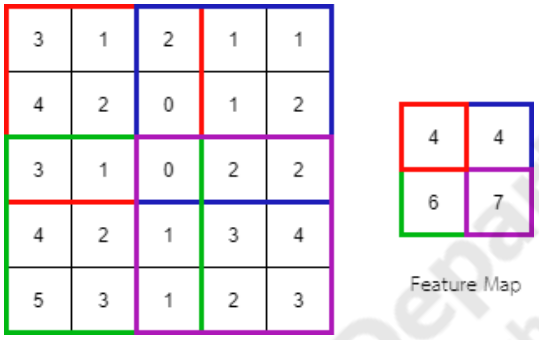
ในส่วนของรูปแบบในการสร้าง Feature Map สามารถกำหนดได้ด้วยรูปแบบของ Filter เพื่อให้สอดคล้องกับจุดประสงค์ในการใช้งาน CNN ได้ ตัวอย่างการใช้รูปแบบต่าง ๆ ของตัวกรอง ดังภาพประกอบที่ 2.6



ภาพประกอบที่ 2.6 ตัวอย่างและผลลัพธ์ของตัวกรองแต่ละรูปแบบ
ที่มา: https://medium.com/@RaghavPrabhu/understanding-of-convolutional-neural-network-cnn-deep-learning-99760835f148

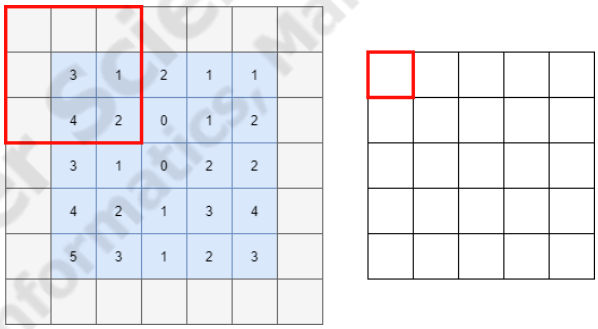
การกำหนดรูปแบบการดำเนินการที่ถูกใช้งานอย่างกว้างขวาง มี 3 รูปแบบ คือ Stride, Padding และ ReLU โดยแต่ละส่วนจะมีการทำงานดังต่อไปนี้

- Stride เป็นตัวกำหนดการเลื่อนของพิกเซล ถ้า Stride มีค่าเป็น 1 ก็จะขยับ 1 พิกเซลหลังการคำนวณในแต่ละรอบ หาก Strid มีค่าเป็น 2 Filter จะขยับไป 2 พิกเซลหลังการคำนวณ แต่เมื่อทำการกำหนดค่า Stride มากขึ้น Feature Map ก็จะมีขนาดเล็กลง



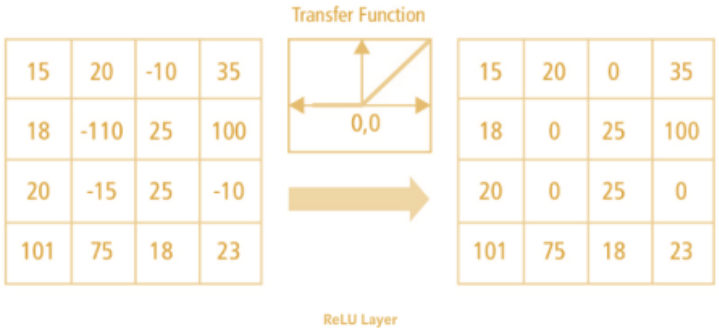
ภาพประกอบที่ 2.7 ตัวอย่าง Stride (2,2)

- Padding เป็นการสร้างพื้นที่รอบๆ Image Matrix และกำหนดค่าให้เป็น 0 (Zero-padding) เพื่อให้ Feature Map มีขนาดพอดีกับภาพเดิม



ภาพประกอบที่ 2.8 ตัวอย่างการทำ Padding

- ReLU หรือ การแก้ไขหน่วยที่ไม่เป็นเชิงเส้นให้เป็นเชิงเส้น (Rectified Linear Unit for a non-linear operation, ReLU)



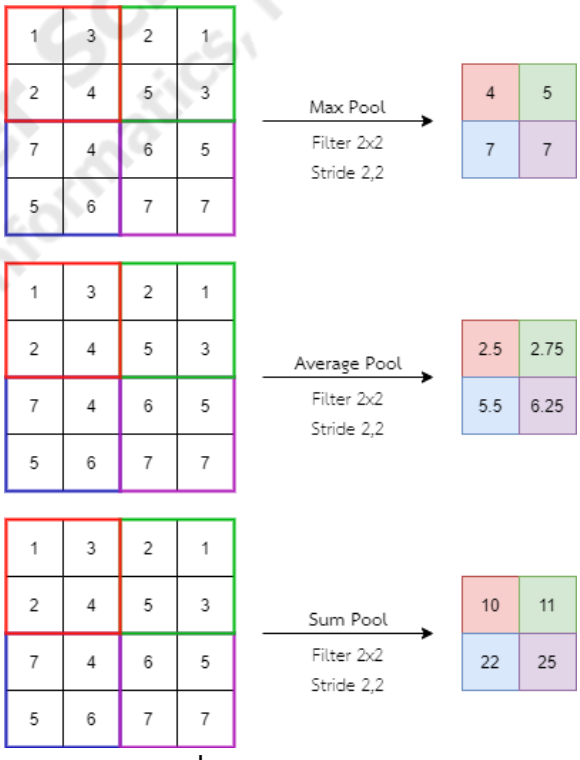
ภาพประกอบที่ 2.9 ตัวอย่างการทำ ReLU

2. Pooling Layer คือชั้นที่ทำหน้าทีในการลดจำนวนของพารามิเตอร์ในกรณีที่ภาพมีขนาดใหญ่มากเกินไป การทำ Pooling หรือที่เรียกว่า Sub-sampling หรือ Down-sampling เป็นกระบวนการลดมิติของ Feature Map ลง แต่ยังคงข้อมูลสำคัญเอาไว้

ภาพประกอบที่ 2.10 ตัวอย่างภาพที่ถูกทำ Pooling
ที่มา: https://www.spacewu.com/posts/s3pool/

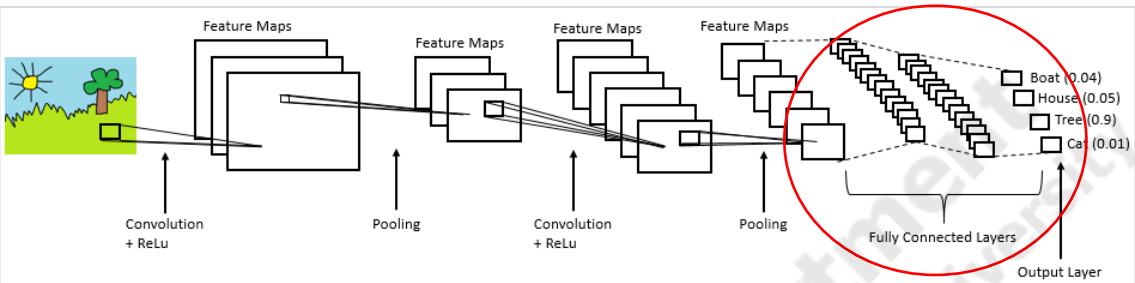
ในการทำ Pooling มีอยู่ 3 รูปแบบ ดังนี้

- Max Pooling เป็นการกำหนดขนาด Pool แล้วนำไปทาบกับ Feature Map เพื่อดึงค่าที่มากที่สุดในแต่ละรอบ แล้วทำการเก็บไว้ใน Feature Map ชุดใหม่
- Average Pooling เป็นการหาค่าเฉลี่ยของแต่ละ Pool
- Sum Pooling เป็นการหาผลรวมของแต่ละ Pool



ภาพประกอบที่ 2.11 ตัวอย่างการทำ Pooling

3. Fully Connected Layer โดยในชั้นนี้จะเป็นการรวบรวม Feature Map ทั้งหมดที่ถูกสร้างขึ้น นำมารวมกันเพื่อทำการสร้างโมเดล (Model) เพื่อใช้ในการคัดแยกและจำแนกรูปแบบข้อมูลต่าง ๆ แล้วจึงทำการใช้ Activation Function ในการจำแนกประเภทของวัตถุต่าง ๆ จากโมเดลที่สร้างขึ้น ตัวอย่างของฟังก์ชันในการจำแนก เช่น SoftMax



ภาพประกอบที่ 2.12 ชุด Feature Maps ใน Fully Connected Layer
ที่มา: https://miro.medium.com/max/700/1*4GLv7_4BbKXnpc6BRb0Aew.png

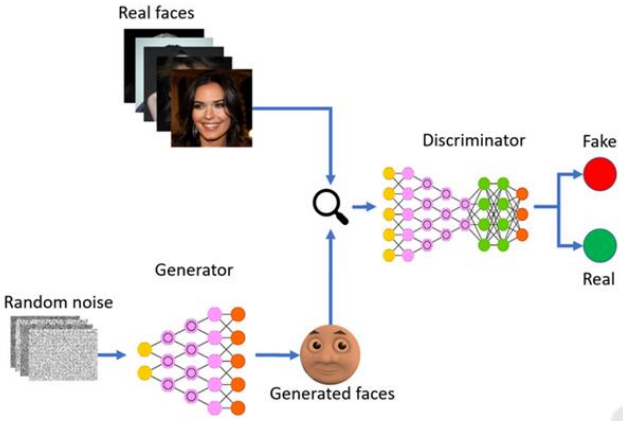
SoftMax [6] เป็นฟังก์ชันในการแปลงค่าน้ำหนักด้วยฟังก์ชันที่กำหนด โดยการนำค่าตัวเลขที่ได้จากช่องผลลัพธ์ทุกช่องไปรวมกัน ซึ่งค่าที่ได้ออกมาจะอยู่ที่ระหว่าง 0 กับ 1 ใช้เป็นตัวกำหนดความเป็นไปได้ของข้อมูล สมการคือ

$$\sigma(\vec{z})i = \frac{e^{z_j}}{\sum_{j=1}^{K} e^{z_j}} \tag{2.1}$$

2.1.5 โครงข่ายประสาทแบบจำลองและโต้แย้ง

โครงข่ายประสาทแบบจำลองและโต้แย้ง (Generative Adversarial Networks, GAN) [7] เทคนิคนี้จะถูกใช้ในการสร้างและวิเคราะห์ภาพจำลองว่ามีความใกล้เคียงภาพใบหน้าคนจริงมากน้อยเพียงใด ซึ่ง GAN เป็นโมเดลที่ใช้อัลกอริทึม Deep Learning เช่น CNN โดยใช้เทคนิคการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) โดยมีรูปแบบการเรียนรู้และผลลัพธ์ของข้อมูลที่ต้องการจากข้อมูลที่ถูกป้อนเข้ามาได้ด้วยตนเอง โมเดล GAN มักถูกใช้ในการจำลองข้อมูลต่าง ๆ ขึ้นมาใหม่ ซึ่งโมเดลนี้สามารถจำลองข้อมูลออกมาได้ใกล้เคียงกับความเป็นจริง (Reality) อย่างมาก

GAN ใช้เทคนิคการเรียนรู้โดยมีโมเดลอยู่สองตัว ตัวแรกคือโมเดลการจำลอง (Generative Model) ที่ใช้สำหรับการจำลองตัวอย่างรูปแบบใหม่ และตัวที่สองคือโมเดลการคัดแยก (Discriminator Model) จะเป็นตัวระบุว่าตัวอย่างที่ถูกจำลองขึ้นมา เป็นภาพจริง (ภาพที่ถูกป้อนเข้ามา) หรือเป็นภาพที่ถูกสร้างขึ้น โมเดลสองตัวนี้จะทำการสร้างและหลอกกันไปเรื่อย ๆ จนกระทั่งโมเดลการจำลองสามารถหลอกโมเดลการคัดแยกได้อย่างมีประสิทธิภาพในระดับที่ต้องการ



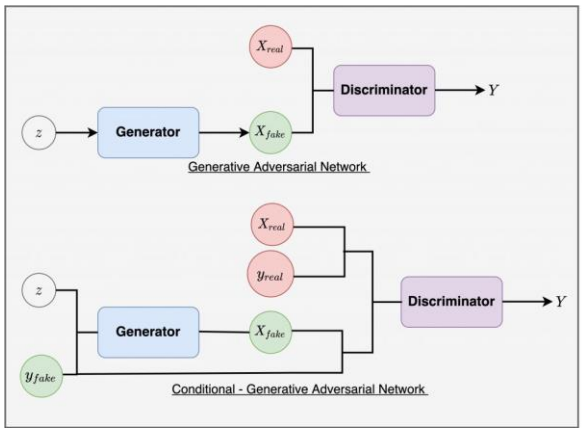
ภาพประกอบที่ 2.13 โครงสร้างการทำงานของ GAN
ที่มา: https://www.quora.com/What-is-the-difference-between-CNNs-and-GANs

2.1.6 โครงข่ายประสาทแบบจำลองและโต้แย้งแบบมีเงื่อนไข

โครงข่างประสาทแบบจำลองและโต้งแย้งแบบมีเงื่อนไข (Conditional GAN) [8] โดยการนำสถาปัตยกรรม GAN มาปรับปรุงโดยการเพิ่มพารามิเตอร์ y ในส่วนของ Generator (ก่อนหน้านี้มีแค่ภาพจุดรบกวน หรือ noise แทนด้วย z) ซึ่งพารามิเตอร์นี้จะเป็นคำนิยาม (Conditioning label) สำหรับการสร้างภาพจำลอง โดยภาพจำลอง $G(z,y) = x|y$ (จำลองภาพ x โดยที่ x มีเงื่อนไขการสร้างจาก y) ซึ่งภาพจำลองนี้สร้างขึ้นเพื่อหลอก Discriminator โดยจำลองให้ใกล้เคียงกับภาพจริงที่สุดจากคำนิยามที่ถูกป้อนเข้ามา

ในการเรียนรู้แต่ละรอบ Discriminator จะทำการเรียนรู้โดยใช้ทั้งภาพจริงและภาพจำลอง และมีคำนิยามของภาพเป็นตัวตรวจสอบค่าความน่าจะเป็นของภาพ เป้าหมายของ Discriminator คือ ยอมรับภาพจริงและคู่คำนิยามทั้งหมด ปฏิเสธภาพจำลองคู่คำนิยามทั้งหมด และปฏิเสธภาพจำลองที่ไม่ตรงกับคู่คำนิยามทั้งหมด โดยวิธีการคำนวณ Validation Loss จะมีหลักการดังนี้

1. หากภาพคือภาพ 1 แต่คำนิยามเป็น 2 ถึงแม้จะเป็นภาพจริงหรือภาพจำลอง ผลลัพธ์ควรถูกปฏิเสธเนื่องจากภาพไม่ตรงกับคำนิยาม
2. ปฏิเสธภาพจำลองทุกภาพถึงแม้ภาพและคำนิยามจะตรงกัน



ภาพประกอบที่ 2.14 เปรียบเทียบสถาปัตยกรรมของ GAN (บน) และ CGAN (ล่าง)
ที่มา: https://learnopencv.com/conditional-gan-cgan-in-pytorch-and-tensorflow/

2.2 ระบบงานที่เกี่ยวข้อง

งานวิจัยของ Rutgers University, New Jersey “High-Quality Facial Photo-Sketch Synthesis Using Multi-Adversarial Networks” [9] เป็นงานวิจัยที่ได้ทำการจำลองใบหน้าคนจากภาพสเกตช์ที่มีรายละเอียดมาก ใช้เทคนิคการเรียนรู้แบบ Novel Synthesis Network ที่เรียกว่า Photo-Sketch Synthesis โดยใช้วิธีการ Multi-Adversarial Network, (PS^2-MAN) ในการสร้างโมเดล เป็นวิธีที่ทำการจำลองภาพไล่ระดับความละเอียดต่ำขึ้นไปเป็นภาพความละเอียดสูง เพื่อให้สามารถสร้างภาพจำลองใบหน้าได้ใกล้เคียงกับใบหน้าจริง และมีรายละเอียดส่วนเกิน (Artifacts) ที่น้อย

งานวิจัยของ The University of Hong Kong, Baidu Research “Semi-Supervised Learning for Face Sketch Synthesis in the Wild” [10] เป็นงานวิจัยที่ทำการจำลองใบหน้าจากภาพสเกตช์ โดยการสร้างแผนภาพจำลองตำแหน่งขององค์ประกอบต่าง ๆ บนใบหน้า ทั้งจากภาพจริงและภาพสเกตช์ และทำการเปรียบเทียบกันโดนใช้วิธีการ composition-aided GAN, (CA-GAN) ในการสร้างโมเดลเพื่อใช้ในการจำลองภาพใบหน้าเสมือนจริง

งานวิจัยร่วมของ SHU-YU CHEN, WANCHAO SU, LIN GAO, SHIHONG XIA, HONGBO FU “DeepFaceDrawing : Deep Generation of Face Images from Sketches” [11] ที่ได้ทำการสร้างแอปพลิเคชันที่สามารถสร้างใบหน้าคนจากภาพสเกตช์ที่มีรายละเอียดน้อย โดยใช้ภาพและภาพสเกตช์ที่มีเพียง 900 คู่ แต่ได้ทำการปรับแต่งภาพให้มีลักษณะที่บิดงอ สร้างออกมาได้ทั้งหมด 17,000 คู่ และใช้เทคนิคการเรียนรู้แบบ Conditional GAN, (C-GAN) ในการสร้างโมเดลการพยากรณ์ใบหน้าที่ใกล้เคียงกับใบหน้าจริงมากที่สุด โดยใช้ภาพสเกตช์ที่มีรายละเอียดไม่มาก หรือใช้ภาพที่มีเพียงไม่กี่เส้นที่ใช้กำหนดเพียงลักษณะโครงหน้าก็สามารถสร้างภาพใบหน้าที่สมจริงออกมาได้

ตารางที่ 2.1 ตารางเปรียบเทียบระบบงานที่เกี่ยวข้อง

ระบบงาน / ฟังก์ชันการทำงาน	High-Quality Facial Photo-Sketch Synthesis	Semi-Supervised Learning for Face Sketch Synthesis	DeepFaceDrawing	Sketch-to-Face Generation
สร้างภาพใบหน้าจากภาพสเกตช์	/		/	/
กำหนดเพศ			/	/
กำหนดลักษณะใบหน้าที่ต้องการ				/
คัดแยกองค์ประกอบบนใบหน้า		/	/	/
สร้างภาพสเกตช์จากภาพใบหน้า	/	/		