การศึกษาเปรียบเทียบแบบจำลองแบบ Transformer สำหรับการตรวจจับภาพปลอม

COMPARATIVE STUDY ON TRANSFORMER-BASED MODEL FOR FAKE IMAGE DETECTION

บุษกร เดชาพงษ์

บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ

2567

การศึกษาเปรียบเทียบแบบจำลองแบบ Transformer สำหรับการตรวจจับภาพปลอม

บุษกร เดชาพงษ์

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล
คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
ปีการศึกษา 2567

COMPARATIVE STUDY ON TRANSFORMER-BASED MODEL FOR FAKE IMAGE DETECTION

BOOTSAGORN DECHAPHONG

An Master's Project Submitted in Partial Fulfillment of the Requirements

for the Degree of MASTER OF SCIENCE

(Data Science)

Faculty of Science, Srinakharinwirot University

2024

สารนิพนธ์

เรื่อง

การศึกษาเปรียบเทียบแบบจำลองแบบ Transformer สำหรับการตรวจจับภาพปลอม

ของ

บุษกร เดชาพงษ์

ได้รับอนุมัติจากบัณฑิตวิทยาลัยให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล

ของมหาวิทยาลัยศรีนครินทรวิโรฒ

(รองศาสตราจารย์ นายแพทย์ฉัตรชัย เอกปัญญาสกุล)

คณบดีบัณฑิตวิทยาลัย

คณะกรรมการสอบปากเปล่าสารนิพนธ์

..... ที่ปรึกษาหลัก

(ผู้ช่วยศาสตราจารย์ ดร.รัตน์ชัยนันท์ ธรรมสุจริต)

..... ประธาน

(ผู้ช่วยศาสตราจารย์ ดร.ศิฟาณี นุชิตประสิทธิ์ชัย)

..... กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.วีรยุทธ เจริญเรืองกิจ)

ชื่อเรื่อง	การศึกษาเปรียบเทียบแบบจำลองแบบ Transformer สำหรับการตรวจจับภาพปลอม
ผู้วิจัย	บุษกร เดชาพงษ์
ปริญญา	วิทยาศาสตรมหาบัณฑิต
ปีการศึกษา	2567
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร. รัตน์ชัยนันท์ ธรรมสุจริต

เทคโนโลยี Deepfake มีการพัฒนาอย่างต่อเนื่อง ส่งผลให้เกิดความเสี่ยงในการบิดเบือนข้อมูลทางดิจิทัลและการเผยแพร่ข้อมูลข่าวปลอมออกสู่สาธารณะ งานวิจัยนี้มีวัตถุประสงค์เพื่อ 1) สร้างแบบจำลองที่สามารถตรวจจับและจำแนกภาพที่สร้างโดย Deepfake จากภาพจริง 2) เปรียบเทียบประสิทธิภาพของโครงข่ายประสาทเทียมแบบ Transformer ได้แก่ Vision Transformer (ViT), Data-efficient Image Transformer (DeiT) และ Swin Transformer ในการจำแนกภาพ Deepfake และภาพจริง ซึ่งงานวิจัยนี้ใช้ชุดข้อมูล “140k Real and Fake Faces” จาก Kaggle โดยสุ่มเลือกข้อมูล 10% ของชุดข้อมูลทั้งหมด คิดเป็น 14,000 ภาพ ซึ่งประกอบด้วยภาพจริง 7,000 ภาพ และภาพปลอม 7,000 ภาพ จากนั้นแบ่งข้อมูลออกเป็นชุดข้อมูลฝึก, ชุดข้อมูลตรวจสอบและชุดข้อมูลทดสอบ ก่อนดำเนินการฝึกแบบจำลอง ข้อมูลทั้งหมดถูกปรับขนาดภาพเป็น 224×224 พิกเซล และใช้เทคนิคการเพิ่มความหลากหลายของข้อมูล (Data Augmentation) เพื่อลดปัญหาการเกิด overfitting นอกจากนี้ชุดข้อมูลฝึกและชุดข้อมูลตรวจสอบได้ผ่านกระบวนการฝึกฝนกับ pre-trained models ก่อนที่จะเริ่มการฝึกแบบจำลองงานวิจัยนี้ประเมินประสิทธิภาพของแบบจำลองผ่านตัวชี้วัด ได้แก่ ความแม่นยำ (Accuracy), ความถูกต้อง (Precision), การเรียกคืน (Recall) และค่า F1-score ผลการทดลองพบว่า Swin Transformer มีค่าความแม่นยำสูงสุดที่ 0.9915 รองลงมาคือ ViT เท่ากับ 0.9720 และ DeiT เท่ากับ 0.9525 แสดงให้เห็นว่า Swin Transformer สามารถเรียนรู้คุณลักษณะของภาพและจำแนกภาพจริงจากภาพปลอมได้อย่างมีประสิทธิภาพมากที่สุด ทั้งนี้ งานวิจัยในอนาคตควรพิจารณาการพัฒนาแบบจำลองเพิ่มเติม เช่น การปรับปรุงคุณภาพของภาพก่อนการทดสอบ การเพิ่มขนาดและความหลากหลายของชุดข้อมูล และการปรับค่าพารามิเตอร์ของแบบจำลองให้เหมาะสมยิ่งขึ้น เพื่อเพิ่มประสิทธิภาพในการจำแนกภาพ Deepfake ได้อย่างแม่นยำยิ่งขึ้น

คำสำคัญ : ภาพปลอม, การตรวจจับภาพปลอม, โครงข่ายประสาทเทียมแบบทรานฟอร์เมอร์, การเรียนรู้เชิงลึก

Title	COMPARATIVE STUDY ON TRANSFORMER-BASED MODEL FOR FAKE IMAGE DETECTION
Author	BOOTSAGORN DECHAPHONG
Degree	MASTER OF SCIENCE
Academic Year	2024
Thesis Advisor	Assistant Professor Dr. Ratchainant Thammasudjarit

Deepfake technology is advancing quickly, increasing the risk of digital misinformation and fake news. This study has two goals: (1) to develop a model that can detect and differentiate Deepfake-generated images from real ones, and (2) to compare the performance of Transformer-based neural networks—specifically Vision Transformer (ViT), Data-efficient Image Transformer (DeiT), and Swin Transformer—for this classification task. We used 10% of the “140k Real and Fake Faces” dataset from Kaggle, consisting of 14,000 images (7,000 real and 7,000 fake), divided into training, validation, and test sets. Before training, all images were resized to 224×224 pixels, and data augmentation was applied to reduce overfitting. We also used pre-trained models to initialize the training process. Performance was measured using accuracy, precision, recall, and F1-score. Results showed that Swin Transformer achieved the highest accuracy at 0.9915, followed by ViT (0.9720) and DeiT (0.9525), confirming Swin Transformer’s effectiveness in distinguishing real from fake images. Future research should focus on improving image quality, expanding the dataset, and fine-tuning model parameters to further enhance deepfake detection.

Keyword : Deepfake, Fake Image Detection, Transformer Neural Network, Deep Learning

กิตติกรรมประกาศ

การจัดทำสารนิพนธ์เล่มนี้สำเร็จลุล่วงไปได้ด้วยดีจากการสนับสนุนและคำแนะนำที่มีคุณค่าจาก ผู้ช่วยศาสตราจารย์ ดร.รัตน์ชัยนันท์ ธรรมสุจริต อาจารย์ที่ปรึกษาของผู้วิจัย ซึ่งได้ให้ความรู้ ความช่วยเหลือ และแนะแนวทางในการทำวิจัย รวมถึงชี้แนะข้อบกพร่องต่างๆ ที่ช่วยให้การวิจัยในครั้งนี้สำเร็จลุล่วงไปได้ ผู้วิจัยขอกราบขอบพระคุณอาจารย์ด้วยความเคารพอย่างสูง ณ โอกาสนี้

ขอขอบคุณประธานและคณะกรรมการการสอบสารนิพนธ์ทุกท่านที่ให้ความอนุเคราะห์เป็นกรรมการสอบ พร้อมทั้งให้คำแนะนำเพื่อปรับปรุงและชี้แนะแนวทางในการทำงานวิจัยจนเป็นผลสำเร็จ

ขอขอบคุณบัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒที่ให้การสนับสนุนในการนำเสนอผลงานวิจัย

สุดท้ายนี้ ขอบคุณครอบครัว และ เพื่อนๆ ทุกท่านที่คอยให้กำลังใจและสนับสนุนตลอดระยะเวลาในการทำงานวิจัยครั้งนี้

บุษกร เดชาพงษ์

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาของงานวิจัย

ในปัจจุบันการสร้างภาพด้วยปัญญาประดิษฐ์ (AI) ได้รับการพัฒนาอย่างก้าวกระโดด จนสามารถสร้างภาพที่มีความสมจริงในเวลาเพียงไม่กี่วินาที (1) เทคโนโลยีที่ใช้ในการสร้างภาพเหล่านี้มีความก้าวหน้าอย่างรวดเร็วในช่วงไม่กี่ปีที่ผ่านมา ทำให้ AI ถูกนำมาใช้สร้างภาพอย่างแพร่หลายในหลากหลายวงการ ทั้งในเชิงศิลปะ ธุรกิจ และอุตสาหกรรมต่าง ๆ ข้อดีของการสร้างภาพด้วย AI นั้นมีมากมาย เช่น ช่วยลดเวลาในการผลิต ลดต้นทุน และลดการพึ่งพาทรัพยากรมนุษย์ นอกจากนี้ยังเปิดโอกาสให้ผู้ที่ไม่มีทักษะด้านศิลปะสามารถสร้างสรรค์ผลงานได้เช่นกัน อย่างไรก็ตาม เทคโนโลยีเหล่านี้ก็มีข้อเสียที่สำคัญ เช่น การนำภาพที่สร้างจาก AI มาใช้แทนที่ผลงานของศิลปิน ซึ่งอาจส่งผลกระทบต่อรายได้และโอกาสในการทำงานของศิลปิน อีกทั้งภาพที่สร้างโดย AI มักขาดความลึกซึ้งทางอารมณ์และความคิดสร้างสรรค์ ซึ่งเป็นจุดเด่นของงานศิลปะที่สร้างโดยมนุษย์ นอกจากนี้ ยังมีปัญหาด้านลิขสิทธิ์ เมื่อภาพที่สร้างจาก AI ละเมิดสิทธิ์ของศิลปินโดยไม่ตั้งใจ รวมถึงการนำภาพเหล่านี้ไปใช้ในทางที่ผิด เช่น การสร้างข่าวปลอม (Fake News) หรือ การสร้างภาพปลอม (Deepfake) ซึ่งอาจทำลายชื่อเสียงหรือก่อให้เกิดความเสียหายต่อบุคคลและสังคม การพัฒนาที่รวดเร็วของ AI ในการสร้างภาพทำให้การตรวจจับและแยกแยะภาพสังเคราะห์จากภาพจริงยากขึ้นเรื่อย ๆ ดังนั้น การวิจัยและพัฒนาเทคโนโลยีที่ใช้ในการตรวจจับภาพที่สร้างจากปัญญาประดิษฐ์จึงมีความสำคัญอย่างยิ่งในยุคปัจจุบัน เพื่อลดผลกระทบที่อาจเกิดขึ้นจากการใช้เทคโนโลยีนี้ในทางที่ไม่เหมาะสม

การสร้างภาพปลอม (Deepfake) คือเทคโนโลยีที่ใช้ในการสร้างภาพและวิดีโอด้วยปัญญาประดิษฐ์ โดยชื่อ "Deepfake" มาจากการรวมกันของคำว่า Deep Learning และ Fake ซึ่งหมายถึงการดัดแปลงภาพ เสียง หรือวิดีโอให้ดูเหมือนจริงที่สุดด้วยเทคโนโลยีการเรียนรู้เชิงลึก ตัวอย่างการใช้งานของ Deepfake รวมถึงการเปลี่ยนแปลงใบหน้าของบุคคลในภาพหรือวิดีโอ (Face Swapping) ซึ่งสามารถแบ่งออกเป็นสองประเภทหลัก คือ Lip-sync Deepfake วิดีโอที่มีการแก้ไขให้การเคลื่อนไหวของปากผู้พูดตรงกับเสียงที่ถูกใส่เข้าไป และ Puppet-master Deepfake วิดีโอที่มีการควบคุมการเคลื่อนไหวของบุคคลเป้าหมายให้สอดคล้องกับการแสดงออกทางใบหน้า ตา และศีรษะตามที่ผู้ควบคุมต้องการ ในปัจจุบัน การสร้าง Deepfake นิยมใช้เทคนิคการเรียนรู้เชิงลึก เช่น Autoencoders และ Generative Adversarial Networks (GANs) ซึ่ง

นำไปใช้ในด้านการประมวลผลภาพอย่างแพร่หลาย โดยการสร้าง Deepfake สามารถนำไปใช้ในทางบวก เช่น การผลิตวิชวลเอฟเฟกต์ การสร้างภาพอวตารดิจิทัล ฟิลเตอร์ในแอพพลิเคชันอย่าง Snapchat หรือการปรับปรุงการสร้างภาพยนตร์ อย่างไรก็ตาม ตัวอย่างการใช้ Deepfake ในทางลบก็มักจะมีจำนวนมากกว่า เช่น การเปลี่ยนใบหน้าของบุคคลธรรมดา ดารา หรือคนดังไปเป็นภาพในภาพที่มีเนื้อหาลามกอนาจาร การใช้ Deepfake ในการสร้างข่าวปลอมทางการเมืองหรือการทหาร รวมถึงการปลอมแปลงเสียงเพื่อหลอกลวงให้ผู้เสียหายโอนเงิน เป็นต้น ดังนั้น การนำ Deepfake ไปใช้ในเชิงลบจึงมีผลกระทบที่มากต่อสังคมและมนุษย์โดยรวม

ในงานวิจัยนี้มีการศึกษาเพื่อเปรียบเทียบเทคนิคโครงข่ายประสาทเทียมแบบ Transformer สำหรับการตรวจจับ Deepfake โดยมุ่งหวังที่จะพัฒนาความสามารถในการแยกแยะระหว่างภาพที่สร้างขึ้นด้วยเทคโนโลยี Deepfake และภาพจริงอย่างมีความแม่นยำและมีประสิทธิภาพ

1.2 ความมุ่งหมายของงานวิจัย

ในการวิจัยครั้งนี้ ผู้วิจัยมีความมุ่งหมายดังต่อไปนี้

1. สร้างแบบจำลองเพื่อตรวจจับและจำแนกภาพที่ถูกสร้างขึ้นโดยเทคโนโลยี Deepfake จากภาพจริง

2. เพื่อเปรียบเทียบประสิทธิภาพของเทคนิคโครงข่ายประสาทเทียมแบบ Transformer ระหว่างแบบจำลอง Vision Transformer (ViT), Data-efficient Image Transformer (DeiT) และ Swin Transformer ในการจำแนกระหว่างภาพ Deepfake และภาพจริง

1.3 ความสำคัญของการวิจัย

1. งานวิจัยนี้มีความสำคัญในการช่วยลดความเสี่ยงจากการเผยแพร่ข้อมูลที่ถูกบิดเบือนผ่านเทคโนโลยี Deepfake โดยการพัฒนาวิธีการตรวจจับภาพ Deepfake จะทำให้ผู้ใช้สามารถยืนยันได้ว่าข้อมูลที่ได้รับนั้นเป็นภาพจริงหรือภาพปลอมซึ่งจะช่วยสร้างความเชื่อมั่นในความถูกต้องของข้อมูลในยุคที่มีการหลอกลวงด้วยสื่อดิจิทัลอย่างแพร่หลายงานวิจัยนี้มีความสำคัญต่อการพัฒนาและปรับปรุงแบบจำลองที่มีความสามารถในการ

2. จำแนกและระบุภาพ Deepfake จากภาพจริงได้อย่างแม่นยำ ซึ่งจะนำไปสู่การสร้างเทคโนโลยีที่มีประสิทธิภาพสูงขึ้นในการตรวจสอบและวิเคราะห์ภาพ Deepfake

1.4 ขอบเขตของงานวิจัย

ประชากรที่ใช้ในงานวิจัย

ประชากรที่ใช้ในงานวิจัยนี้ คือชุดข้อมูลทั้งหมดภาพใบหน้าที่ประกอบ ภาพใบหน้าทั้งหมด 140,000 ภาพ แบ่งเป็นภาพใบหน้าจริงจำนวน 70,000 ภาพ และภาพใบหน้าปลอมจำนวน 70,000 ภาพที่สุ่มเลือกจากชุดข้อมูลใบหน้าปลอมจำนวน 1 ล้านภาพ (ซึ่งถูกสร้างขึ้นโดย StyleGAN) ในงานวิจัยเลือกใช้ข้อมูล 10 % จากข้อมูลทั้งหมด โดยทำการปรับขนาดของภาพทั้งหมดให้เป็นขนาด 224x224 พิกเซล และแบ่งข้อมูลออกเป็นชุดข้อมูลฝึก (Training set), ชุดข้อมูลตรวจสอบ (Validation set) และชุดข้อมูลทดสอบ (Test set)

กลุ่มตัวอย่างที่ใช้ในงานวิจัย

งานวิจัยนี้ใช้ชุดข้อมูลชื่อ “140k Real and Fake Faces” จาก Kaggle (2) ภาพชุดข้อมูลทั้งหมด 140,000 ภาพที่ประกอบด้วย ภาพใบหน้าจริงจำนวน 70,000 ภาพ และ ภาพใบหน้าปลอมที่สร้างโดย deepfake จำนวน 70,000 ภาพ ในงานวิจัยนี้ในเลือกใช้ข้อมูลแค่ 10% ของจำนวนภาพทั้งหมด ใช้ข้อมูลเป็น 14,000 ภาพ แบ่งเป็นภาพใบหน้าจริงจำนวน 7,000 ภาพ และภาพใบหน้าปลอมจำนวน 7,000 ภาพ แบ่งเป็น ข้อมูลชุดข้อมูลฝึก (Training set), ชุดข้อมูลตรวจสอบ (Validation set) และ ชุดข้อมูลทดสอบ (Test set)

ตัวแปรที่ศึกษา

1. ตัวแปรอิสระ แบ่งเป็นดังนี้

1.1 ประเภทของแบบจำลองที่ใช้เทคนิคโครงข่ายแบบ Transformer สำหรับการตรวจจับภาพ deepfake ประกอบด้วยแบบจำลอง 3 ประเภทหลัก คือ ViT, DeiT และ Swin Transformer

2. ตัวแปรตาม แบ่งเป็นดังนี้

2.1 ค่าประสิทธิภาพในการจำแนกภาพปลอม (deepfake) กับภาพจริง ซึ่งประกอบด้วยค่า Accuracy, Precision, Recall และ F1-score

1.5 กรอบแนวคิดในงานวิจัย

กรอบแนวคิดในงานวิจัยนี้มุ่งเน้นการเปรียบเทียบประสิทธิภาพของแบบจำลองที่ใช้เทคนิคโครงข่ายประสาทเทียมแบบ Transformer เพื่อแยกแยะระหว่างภาพ Deepfake และภาพจริง โดยใช้ชุดข้อมูลชื่อ “140k Real and Fake Faces” จาก Kaggle ซึ่งประกอบด้วยภาพใบหน้าจริงจำนวน 70,000 ภาพ และภาพใบหน้าปลอมอีก 70,000 ภาพ ข้อมูลเหล่านี้ถูกแบ่งออกเป็นชุดข้อมูลฝึก (training set), ชุดข้อมูลตรวจสอบ (validation set), และชุดข้อมูลทดสอบ (test set) โดยการฝึกจะใช้ข้อมูล 10% จากชุดข้อมูลทั้งหมด งานวิจัยนี้จะเปรียบเทียบประสิทธิภาพในการจำแนกภาพ Deepfake โดยใช้เทคนิคโครงข่ายประสาทเทียมแบบ Transformer ระหว่างแบบจำลอง ViT , DeiT และ Swin Transformer เพื่อหาค่าที่เหมาะสมที่สุดสำหรับการทำงานของแบบจำลองในการประเมินผลของแต่ละแบบจำลองจะพิจารณาจากตัวชี้วัดต่างๆ เช่น Accuracy, Precision, Recall, และ F1-score เพื่อทำการวิเคราะห์ว่าแบบจำลองใดมีประสิทธิภาพสูงสุดในการตรวจจับภาพ Deepfake

1.6 สมุมติฐานในงานวิจัย

1. การทดสอบด้วยแบบจำลองที่ใช้เทคนิคโครงข่ายประสาทเทียมแบบ Transformer ระหว่างโมเดล ViT, DeiT และ Swin Transformer จะสามารถแยกระหว่างภาพจริงและภาพ deepfake ได้ด้วยค่าความแม่นยำสูงมากกว่า 90 %

2. การเลือกใช้โมเดล Swin Transformer จะมีประสิทธิภาพสูงกว่าโมเดล ViT และ DeiT ในการจำแนกภาพจริง และภาพ Deepfake เนื่องจากการใช้ Window-based Multi-Head Attention ซึ่งช่วยเพิ่มความสามารถในการจับลักษณะเด่นของภาพในพื้นที่เฉพาะ

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในการทำวิจัยครั้งนี้ ผู้วิจัยได้ศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง และได้นำเสนอตามหัวข้อต่อไปนี้

1. ปัญญาประดิษฐ์ที่สร้างสรรค์ (Generative AI)
2. ภาพปลอม (Deepfake)
3. โครงข่ายประสาทเทียมแบบ Transformer
4. วิจัยที่เกี่ยวข้อง

2.1 ปัญญาประดิษฐ์ที่สร้างสรรค์ (Generative AI)

ปัญญาประดิษฐ์ที่สร้างสรรค์ (Generative AI) (3) คือรูปแบบของปัญญาประดิษฐ์ที่สร้างสรรค์ที่สามารถสร้างเนื้อหาใหม่ๆ อัตโนมัติในหลายประเภท เช่น ข้อความ รูปภาพ เพลง และวิดีโอ เป็นต้น โดยไม่จำเป็นต้องมีการแทรกแซงจากมนุษย์ เทคโนโลยีนี้ใช้การเรียนรู้เชิงลึก (Deep Learning) เพื่อวิเคราะห์และเรียนรู้จากข้อมูลที่มีอยู่ จากนั้นจึงสร้างผลลัพธ์ใหม่ๆ ที่ตรงตามความต้องการของผู้ใช้ ซึ่งความสามารถของปัญญาประดิษฐ์ที่สร้างสรรค์ช่วยเพิ่มประสิทธิภาพในการทำงานในหลายวงการ เช่น การตลาดดิจิทัล, วงการเพลง และกราฟิกดีไซน์ เป็นต้น ช่วยให้กระบวนการทำงานเป็นไปได้อย่างสะดวกและรวดเร็วยิ่งขึ้น

2.1 วิธีการทำงานของปัญญาประดิษฐ์ที่สร้างสรรค์

ทำงานโดยใช้โครงข่ายประสาทเทียม (Neural Networks) ซึ่งเป็นหนึ่งในเทคนิคของปัญญาประดิษฐ์ที่ช่วยให้คอมพิวเตอร์สามารถประมวลผลข้อมูลในลักษณะที่เลียนแบบการทำงานของสมองมนุษย์ นอกจากนี้ยังใช้แบบจำลองสถิติ (Statistical Models) ซึ่งสร้างข้อมูลใหม่โดยอิงจากการวิเคราะห์สถิติจากข้อมูลที่มีอยู่ หรือ GAN (Generative Adversarial Networks) ซึ่งมีความสามารถในการสร้างข้อมูลที่มีความสมจริงและเรียนรู้ได้โดยไม่ต้องมีการกำกับ (Unsupervised Learning) ทั้งนี้ทำให้ปัญญาประดิษฐ์มีความยืดหยุ่นและสามารถนำไปใช้ในหลายแง่มุมของชีวิตประจำวันได้อย่างมีประสิทธิภาพ

2.1.1 ประเภทปัญญาประดิษฐ์ที่สร้างสรรค์ ยกตัวอย่าง เช่น

2.1.1.1 การสร้างข้อความจากปัญญาประดิษฐ์ (Text Generation) เป็นปัญญาประดิษฐ์ที่สร้างสรรค์ที่สามารถเข้าใจ สรุป และสร้างข้อความใหม่ด้วยความเป็นธรรมชาติ

มักถูกนำไปใช้ในหลายอุตสาหกรรม เช่น การทำ Content Marketing และการขาย โดยเฉพาะในการเขียนอีเมล การสนทนาในงาน Customer Support การเขียนบทความ และการทำ Note Taking เป็นต้น

2.1.1.2 การสร้างภาพจากปัญญาประดิษฐ์ (Image Generation) คือ เทคโนโลยีที่สามารถสร้างภาพจากข้อความ คีย์เวิร์ด หรือ Prompt ที่ผู้ใช้ป้อนเข้าไป นอกจากนี้ยังสามารถปรับเปลี่ยนรูปแบบ มุมมอง ประเภทของภาพ และขนาดตามความต้องการได้ ซึ่งเป็นประโยชน์อย่างมากสำหรับผู้ที่ต้องการทำงานด้านการออกแบบแต่ไม่มีทักษะในการสร้างกราฟิกเอง

2.1.1.3 การสร้างภาพปลอมจากปัญญาประดิษฐ์ (Deepfake Generation) เป็นการใช้ปัญญาประดิษฐ์ในการสร้างวิดีโอหรือภาพที่ปลอมแปลงขึ้นมาให้ดูเหมือนจริง โดยมีการใช้เทคนิค face-swapping (4) หรือ voice mimicry ซึ่งได้รับความนิยมอย่างแพร่หลายในทางลบ เช่นการสร้างข่าวปลอมและการปลอมแปลงข้อมูล

ปัญญาประดิษฐ์ในแต่ละประเภทถูกนำไปใช้ในหลากหลายวงการ ไม่ว่าจะเป็นด้านการตลาด ด้านการออกแบบสื่อสร้างสรรค์ ด้านการศึกษา หรือ ด้านบันเทิง โดยมีการพัฒนาและต่อยอดอย่างรวดเร็ว

2.2 ภาพปลอม (Deepfake)

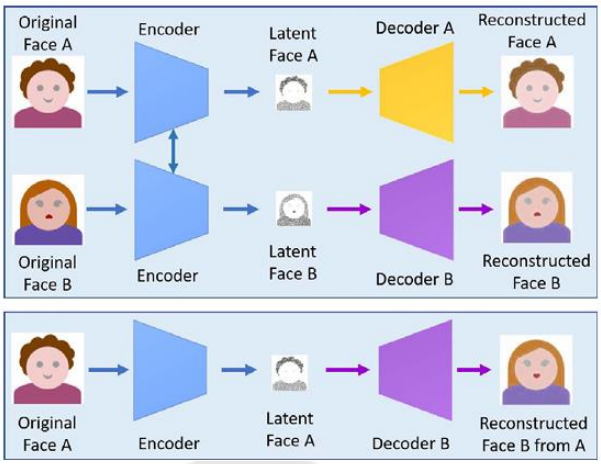
เป็นการผสมผสานระหว่างคำว่า "deep" ซึ่งมาจากเทคโนโลยี deep learning และ "fake" (5) ที่หมายถึงการปลอมแปลง โดย deepfakes ใช้เทคนิค deep learning เพื่อสร้างวิดีโอปลอมที่ดูสมจริงมากขึ้น คาดว่าในอนาคต deepfakes อาจถูกนำไปใช้เป็นเครื่องมือในการเผยแพร่ข้อมูลบิดเบือน ซึ่งอาจทำลายความเชื่อมั่นของสถาบันรัฐ สื่อสารมวลชน และแหล่งข้อมูลต่าง ๆ เนื่องจากประชาชนทั่วไปอาจไม่สามารถแยกแยะวิดีโอจริงกับปลอมได้ ความกังวลเกี่ยวกับ deepfakes เพิ่มมากขึ้นเรื่อย ๆ เนื่องจากการเข้าถึงเทคโนโลยีนี้ทำได้ง่าย และสามารถถูกนำไปใช้ในทางที่ผิด เช่น การหลอกลวงหรือก่ออาชญากรรมทางไซเบอร์ นอกจากนี้ deepfakes ยังเป็นภัยคุกคามต่อประชาธิปไตย เพราะสามารถเผยแพร่ข้อมูลปลอมได้อย่างรวดเร็วและกว้างขวาง ซึ่งข่าวปลอม (Fake News) (6) ได้กลายเป็นปัญหาที่สำคัญที่ส่งผลกระทบต่อสังคมและการแลกเปลี่ยนความคิดเห็นของประชาชน โดยข่าวปลอมหมายถึง เนื้อหาที่ถูกสร้างขึ้นโดยไม่มีมูลความจริง ซึ่งมีจุดประสงค์เพื่อบิดเบือนข้อมูลและโน้มน้าวความคิดเห็นของประชาชน (7) สื่อสังคมออนไลน์ (Social Media) เป็นช่องทางสำคัญในการแพร่กระจายข้อมูลเท็จ ซึ่งสามารถส่งผลกระทบต่อผู้ใช้จำนวนมากในเวลาอันสั้น ปัจจุบัน โลกกำลังก้าวเข้าสู่ ยุคหลังความจริง (Post-

Truth Era) ซึ่งเป็นช่วงเวลาที่ข้อมูลเท็จถูกเผยแพร่อย่างรวดเร็วผ่านสื่อดิจิทัล สงครามข้อมูล (Information Warfare) ได้กลายเป็นเครื่องมือที่ผู้ไม่หวังดีใช้ในการควบคุมความคิดเห็นของประชาชนผ่านการสร้างข่าวปลอมและแคมเปญบิดเบือนข้อมูล ตัวอย่างเช่น การนำไปใช้เพื่อใส่ร้ายบุคคลสำคัญ หรือ แบล็กเมลเพื่อผลประโยชน์ทางการเงิน รวมถึงการสร้างความขัดแย้งทางการเมืองหรือศาสนา แม้ว่า deepfakes มีการนำมาใช้ในเชิงบวกบ้าง เช่น การสร้างเสียงพูดให้ผู้ที่สูญเสียความสามารถในการพูด การใช้ในงานบันเทิง หรือลดค่าใช้จ่ายในการถ่ายทำภาพยนตร์ แต่ผลกระทบด้านลบกลับมีมากกว่า เช่น การใช้ภาพปลอมในการเลือกตั้งหรือการปลอมแปลงข้อมูลทางการทหารด้วยความก้าวหน้าของอัลกอริทึม deep learning ในปัจจุบัน การสร้างเนื้อหาปลอมจึงทำได้ง่ายยิ่งขึ้น แม้จะมีข้อมูลเพียงเล็กน้อย ตัวอย่างเช่น แอปพลิเคชัน Zao ของจีนที่อนุญาตให้ผู้ใช้นำใบหน้าของตนเองไปใส่ในตัวละครจากภาพยนตร์หรือซีรีส์ ดังนั้น การพัฒนาแนวทางในการตรวจสอบและป้องกันข่าวปลอม จึงมีความสำคัญอย่างยิ่ง โดยเฉพาะเมื่อ Deepfake กลายเป็นเครื่องมือที่สามารถสร้างและเผยแพร่เนื้อหาปลอมได้อย่างสมจริง งานวิจัยในปัจจุบันจึงมุ่งเน้นการพัฒนา เทคโนโลยีการตรวจจับและลดผลกระทบจาก Deepfake รวมถึงการสร้างมาตรการในการป้องกันการเผยแพร่ข้อมูลที่ไม่ถูกต้อง ซึ่งจะช่วยรักษาความน่าเชื่อถือของแหล่งข้อมูลในยุคดิจิทัล และลดผลกระทบจากการบิดเบือนข้อมูลในอนาคต

2.2.1 การสร้าง Deepfake

Deepfakes (1)ได้รับความนิยมเนื่องจากคุณภาพของวิดีโอที่ถูกปรับแต่งและการใช้งานที่ง่าย แม้แต่ผู้ที่มีทักษะการใช้งานคอมพิวเตอร์ในระดับพื้นฐานก็สามารถใช้ได้ โดยการสร้าง Deepfake มักจะพัฒนาโดยใช้เทคนิคการเรียนรู้เชิงลึก (Deep Learning) ซึ่งเป็นเทคโนโลยีที่มีความสามารถในการจัดการกับข้อมูลที่ซับซ้อนและมีมิติสูง ส่วนใหญ่โมเดลที่ใช้ในการสร้าง Deepfake คือ

2.2.1.1 Autoencoders ที่ใช้สำหรับการลดมิติของข้อมูลและการบีบอัดภาพ แอปพลิเคชัน Deepfake ตัวแรกที่เปิดตัวคือ FakeApp (8) โดยใช้โครงสร้างการจับคู่ระหว่าง autoencoder และ decoder ซึ่งช่วยให้สามารถสลับใบหน้าระหว่างภาพต้นทางและภาพเป้าหมายได้ โดยใช้ encoder หนึ่งตัวร่วมกันระหว่างโมเดลทั้งสองชุด ซึ่งช่วยให้ระบบเรียนรู้ลักษณะใบหน้าที่คล้ายกัน เช่น ตำแหน่งของดวงตา จมูก และปาก ทำให้สามารถสลับใบหน้าได้ง่ายขึ้น



ภาพประกอบ 1 ภาพตัวอย่างการทำงาน Autoencoders

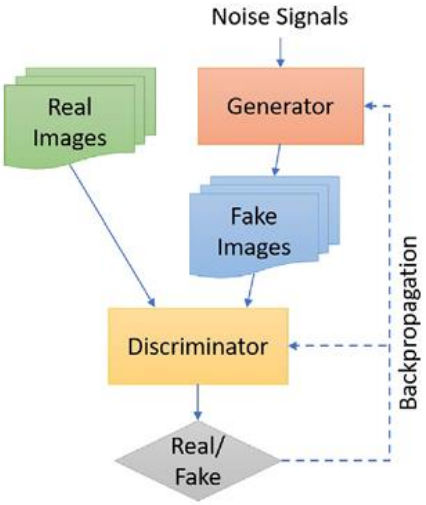
ที่มา: (1)

จากภาพประกอบ 1 คือ โมเดลการสร้าง deepfake โดยใช้คู่ encoder–decoder สองคู่ ทั้งสองเครือข่ายใช้ encoder ตัวเดียวกัน แต่ใช้ decoder ที่แตกต่างกันในกระบวนการฝึก (ดังแสดงในภาพประกอบ 1 ส่วนภาพด้านบน) ภาพใบหน้า A จะถูกแปลงเป็นรูปแบบที่สามารถใช้ร่วมกับ decoder B เพื่อสร้าง deepfake (ดังแสดงในภาพประกอบ 1 ส่วนภาพด้านล่าง) ภาพที่ได้จากการสร้างใหม่ คือใบหน้า B ที่มีรูปปากของใบหน้า A โดยที่ใบหน้า B ดั้งเดิมมีรูปปากเป็นหัวใจกลับด้าน แต่ภาพที่ถูกสร้างขึ้นใหม่จะมีรูปปากเป็นหัวใจปกติ

2.2.1.2 **Generative Adversarial Networks (GAN)** แบบจำลองการเรียนรู้เชิงลึก (deep learning) ประกอบด้วย 2 เครือข่ายหลักคือ

- **Generator** (เครือข่ายผู้สร้าง) มีหน้าที่สร้างข้อมูลใหม่ เช่น ภาพ หรือเสียง ที่พยายามให้เหมือนกับข้อมูลจริงมากที่สุดจากการป้อนสัญญาณรบกวน (noise) เข้ามา
- **Discriminator** (เครือข่ายผู้แยกแยะ) ทำหน้าที่ตรวจสอบว่า ข้อมูลที่ได้รับมานั้นเป็นข้อมูลจริงหรือข้อมูลที่ถูกสร้างขึ้นโดย Generator

ในกระบวนการนี้ Discriminator จะถูกฝึกให้สามารถแยกแยะภาพจริงจากภาพปลอมได้อย่างแม่นยำ ส่วน Generator จะถูกฝึกให้สร้างภาพที่หลอกให้ Discriminator เชื่อว่าเป็นภาพจริง เมื่อทั้งสองเครือข่ายฝึกไปพร้อมกัน Generator จะสามารถสร้างภาพที่สมจริงขึ้นเรื่อย ๆ และ Discriminator ก็จะมีความสามารถในการตรวจจับภาพปลอมได้ดีขึ้น กระบวนการนี้เรียกว่าเป็นการแข่งกันระหว่างสองเครือข่ายเพื่อพัฒนาความสามารถในการสร้างภาพ (9)



ภาพประกอบ 2 โครงสร้างการทำงานของ GAN

ที่มา:(1)

จากภาพประกอบ 2 คือ โครงสร้างของ GAN (Generative Adversarial Network) ประกอบด้วยตัวสร้าง (generator) และตัวจำแนก (discriminator) ซึ่งแต่ละตัวสามารถนำไปใช้ในเครือข่ายประสาทเทียม (neural network) ได้ ระบบทั้งหมดสามารถฝึกได้โดยใช้กระบวนการ backpropagation ซึ่งช่วยให้ทั้งสองเครือข่ายสามารถพัฒนาความสามารถของตนเองได้

ตัวอย่างแบบจำลอง GANs

1) StyleGAN (10) เป็นสถาปัตยกรรมของเครือข่ายประสาทเทียมประเภท Generative Adversarial Network (GAN) ที่ได้รับการพัฒนาโดยบริษัท NVIDIA โดยมีวัตถุประสงค์ในการสังเคราะห์ภาพที่มีความสมจริงสูง ผ่านการปรับปรุงกลไกของเครือข่ายกำเนิด (generator) ให้สามารถควบคุมลักษณะของภาพที่สร้างขึ้นได้อย่างละเอียดและมีประสิทธิภาพมากยิ่งขึ้น โดยสถาปัตยกรรมของ StyleGAN ประกอบด้วยเครือข่ายแปลงเวกเตอร์แฝง (Mapping Network) ซึ่งทำหน้าที่แปลงเวกเตอร์ z จาก latent space ไปเป็นเวกเตอร์ w ใน latent space ใหม่ที่มีคุณสมบัติเหมาะสมในการควบคุมลักษณะภาพมากกว่า จากนั้นเวกเตอร์ w ดังกล่าวจะถูกนำไปใช้ในการปรับค่าการทำงานของเลเยอร์ต่าง ๆ ในเครือข่ายสังเคราะห์ภาพ (Synthesis Network) ผ่านกลไก Adaptive Instance Normalization (AdaIN) ซึ่งทำให้สามารถควบคุม

คุณลักษณะของภาพ เช่น รูปทรงโดยรวม สีพื้นผิว ไปจนถึงรายละเอียดปลีกย่อยได้ในแต่ละระดับความละเอียดของภาพ นอกจากนี้ StyleGAN ยังใช้เทคนิค Progressive Growing ในการเพิ่มขนาดภาพจากความละเอียดต่ำไปสู่ความละเอียดสูงอย่างค่อยเป็นค่อยไป เพื่อเพิ่มเสถียรภาพในการฝึกเครือข่าย ตลอดจนใช้เทคนิคการสุ่มผสมเวกเตอร์แฝง (Mixing Regularization) และการเพิ่มสัญญาณรบกวนแบบ Gaussian ในแต่ละเลเยอร์เพื่อเพิ่มความหลากหลายและความสมจริงของภาพที่ได้ โดยสถาปัตยกรรมนี้สามารถสังเคราะห์ภาพที่มีความสมจริงและคุณภาพสูงได้อย่างมีนัยสำคัญ และได้รับการยอมรับอย่างแพร่หลายในงานวิจัยด้านการสร้างภาพด้วยปัญญาประดิษฐ์

2) GANimation (11) เป็นแบบจำลองที่ออกแบบมาเพื่อสร้างการแสดงออกทางสีหน้าและการเคลื่อนไหวของใบหน้าที่สมจริง โดยได้รับการฝึกอบรมจากชุดข้อมูลที่เกี่ยวข้องกับอารมณ์และสีหน้าของมนุษย์ จึงสามารถนำไปประยุกต์ใช้ในการสร้างวิดีโอ Deepfake ที่มีความสมจริงสูง

3) Face Swapping GAN หรือ FSGAN (12) เป็นแบบจำลองที่ใช้ในการเปลี่ยนแปลงใบหน้าในวิดีโอ สามารถถ่ายทอดการแสดงออกทางสีหน้าของบุคคลหนึ่งไปยังอีกบุคคลหนึ่งได้อย่างมีประสิทธิภาพ ซึ่งทำให้สามารถสร้างสื่อ Deepfake ที่มีความซับซ้อนและสมจริง

4) CycleGAN (13) เป็นแบบจำลองเครือข่ายปัญญาประดิษฐ์เชิงกำเนิด (GAN) ที่สามารถแปลงภาพระหว่างสองโดเมนโดยไม่ต้องใช้ข้อมูลที่จับคู่กันโดยตรง ถูกนำไปใช้ในการเปลี่ยนสไตล์ภาพ ปรับสภาพแวดล้อมในภาพถ่าย และประยุกต์ในงานด้านการแพทย์ เช่น การปรับปรุงภาพถ่ายทางการแพทย์

2.2.2 การตรวจจับ deepfake

กระบวนการตรวจจับเนื้อหาที่ถูกปลอมแปลงหรือแก้ไขโดยใช้เทคโนโลยีปัญญาประดิษฐ์ (AI) โดยเฉพาะอย่างยิ่งในภาพและวิดีโอ ซึ่งเทคโนโลยีที่นิยมใช้ในการสร้าง deepfake คือ Generative Adversarial Networks (GANs) ที่สามารถสร้างภาพใบหน้า เสียง หรือการเคลื่อนไหวที่ดูเหมือนจริงได้ การตรวจจับ deepfake มักใช้การวิเคราะห์ความผิดปกติในเนื้อหา เช่น การตรวจสอบรายละเอียดของภาพหรือวิดีโอ เช่น การกระพริบตาที่ไม่เป็นธรรมชาติ (Detecting Eye Blinking) (14) การวิเคราะห์โครงสร้างเสียง การตรวจสอบความผิดปกติของการเคลื่อนไหวของใบหน้า การตรวจจับ Deepfake ใช้เทคโนโลยีและเทคนิคหลายรูปแบบ ด้าน Machine Learning (ML) และ Deep Learning (15) ใช้อัลกอริทึมในการฝึกโมเดลให้รู้จักความ

แตกต่างระหว่างภาพหรือวิดีโอจริงกับภาพปลอม แบบจำลองเหล่านี้สามารถตรวจจับรายละเอียดที่มนุษย์มองไม่เห็นได้ เช่น โครงสร้างใบหน้าและการเคลื่อนไหว ยกตัวอย่างเช่นแบบจำลองที่ใช้ดังนี้

2.2.2.1 Recurrent Neural Networks (RNNs) และ Long Short-Term Memory (LSTM) ใช้ในการวิเคราะห์ลำดับเวลา โดยเฉพาะในวิดีโอ เพื่อตรวจสอบการเคลื่อนไหวของใบหน้าที่ไม่สอดคล้องกัน เช่น การกระพริบตา หรือการเคลื่อนไหวของปากที่ไม่ตรงกับเสียง เป็นต้น จากงานวิจัยในปี 2018 Guera และ J.Delp (16) ได้ใช้วิธีเสียงและภาพร่วมกันในการตรวจจับความไม่สอดคล้องระหว่างเสียงพูดและการเคลื่อนไหวของริมฝีปาก วิธีนี้ช่วยให้เห็นภาพว่าใบหน้าที่สร้างขึ้นเลียนแบบการเคลื่อนไหวของริมฝีปากได้อย่างไร และสามารถซิงโครไนซ์การเคลื่อนไหวของริมฝีปากกับเสียงพูดได้อย่างแม่นยำ อย่างไรก็ตาม ผู้เขียนวิจัยพบว่า วิธีการที่ใช้การซิงค์ริมฝีปากไม่สามารถตรวจจับความแตกต่างระหว่างการเคลื่อนไหวของริมฝีปากและเสียงพูดได้ นอกจากนี้ยังได้ทดสอบการใช้วิธีการตรวจจับการสลับใบหน้า (face swap detection) พื้นฐานด้วย ต่อมาในปีเดียวกัน Li และ Lyu (17) ได้นำเสนอวิธีใหม่ในการตรวจจับ Deepfake ที่แตกต่างจากวิธีเดิม ๆ โดยใช้ทั้งข้อมูลเสียงและวิดีโอ จากวิดีโอเดียวกันในเวลาเดียวกัน พร้อมทั้งพิจารณาอารมณ์จากทั้งสองข้อมูล วิธีนี้ใช้การเรียนรู้จากข้อมูลและได้พบว่า การใช้ข้อมูลจากหลายแหล่งจะช่วยให้สามารถตรวจจับการแสดงออกทางอารมณ์ที่คล้ายกันได้ เช่น การยกคิ้ว, การขยายตา, โทนเสียง และความดังของเสียง ซึ่งพิจารณาข้อมูลจากทั้งสองแหล่งและการแสดงออกทางอารมณ์ในการจำแนกภาพจริงและปลอมจาก Deepfakes

2.2.2.2 Convolutional Neural Networks (CNNs) เป็นแบบจำลอง deep learning ที่นิยมใช้ในการตรวจจับภาพปลอม โดยสามารถแยกแยะลักษณะเฉพาะของภาพ เช่น ผิวหนังที่ไม่เป็นธรรมชาติหรือการบิดเบือนในภาพ เป็นต้น ซึ่งในปี 2021 Zhou และคณะ (18) ได้พัฒนาโครงสร้างสถาปัตยกรรมโครงข่ายประสาทเทียมแบบคอนโวลูชัน (CNN) โดยใช้ GoogleNet เพื่อตรวจจับใบหน้าที่ถูกแก้ไขในภาพนิ่ง อย่างไรก็ตาม งานวิจัยดังกล่าวไม่ได้เสนอแนวทางสำหรับการตรวจจับการเปลี่ยนแปลงในสตรีมวิดีโอ ในขณะเดียวกัน Afchar และคณะ (19) ได้เสนอแบบจำลอง CNN สองรูปแบบสำหรับการตรวจจับการเลียนแบบใบหน้า โดยหนึ่งในนั้นคือ MesoNet ซึ่งเป็นแบบจำลอง Deep Learning ที่ได้รับความนิยมในการตรวจจับภาพปลอม โดยสามารถแยกลักษณะเฉพาะของภาพ เช่น ผิวหนังที่ไม่เป็นธรรมชาติและการบิดเบือนในภาพได้อย่างมีประสิทธิภาพ MesoNet ถูกออกแบบมาเพื่อจำแนกใบหน้าจริงและใบหน้าปลอมที่สร้างขึ้นโดยเครื่องมือ Face2Face (20) และ DeepFake โดยตรง

2.2.2.3 Generative Adversarial Networks (GANs) แม้ว่าแบบจำลอง GANs จะถูกใช้ในการสร้าง deepfake แต่ก็มีการใช้งานในการตรวจจับด้วย โดยใช้แบบจำลอง discriminative network เพื่อตรวจสอบและแยกแยะความแตกต่างระหว่างภาพจริงกับภาพที่สร้างขึ้น

2.2.2.4 Transformer เป็นเทคโนโลยีได้รับความนิยมอย่างมากในงานด้านการประมวลผลภาพและวิดีโอ เนื่องจากสามารถจับความสัมพันธ์เชิงพื้นที่และเชิงบริบทได้ดี เช่น แบบจำลอง Vision Transformer (ViT), DeiT และ Swin Transformer ที่ใช้ Self-Attention Mechanism เพื่อวิเคราะห์ความสัมพันธ์ในภาพแบบละเอียดและตรวจจับลักษณะเฉพาะของภาพ Deepfake ได้อย่างมีประสิทธิภาพ โดยเฉพาะในกรณีที่ภาพมีความละเอียดสูงหรือความผิดปกติที่ซับซ้อนด้วยศักยภาพอันโดดเด่นของเทคโนโลยี Transformer ในการเรียนรู้และจับความสัมพันธ์เชิงพื้นที่และบริบทได้อย่างแม่นยำ ทำให้ได้รับความสนใจอย่างแพร่หลายในงานด้าน Computer Vision โดยเฉพาะการประมวลผลภาพและวิดีโอ ซึ่งสามารถนำมาใช้เป็นรากฐานสำคัญในการพัฒนาโครงข่ายประสาทเทียมสมัยใหม่ที่มีความสามารถสูงขึ้น โมเดลที่ใช้เทคนิค Transformer มีข้อได้เปรียบเหนือสถาปัตยกรรมแบบดั้งเดิม เช่น Convolutional Neural Networks (CNNs) เนื่องจากสามารถเรียนรู้ข้อมูลเชิงลึกที่ซับซ้อน และวิเคราะห์โครงสร้างของข้อมูลภาพได้อย่างมีประสิทธิภาพมากขึ้น จากคุณสมบัติที่โดดเด่นดังกล่าวเทคโนโลยี Transformer และสถาปัตยกรรมที่เกี่ยวข้อง ได้ถูกนำมาใช้ในงานวิจัยหลายแขนง รวมถึงการตรวจจับ Deepfake ซึ่งเป็นหนึ่งในประเด็นสำคัญของความมั่นคงทางดิจิทัลในปัจจุบัน การนำเทคนิคทรานฟอร์เมอร์ มาใช้ในงานด้านนี้ช่วยเพิ่มขีดความสามารถของระบบวิเคราะห์ภาพและวิดีโอ ทำให้สามารถตรวจจับความผิดปกติของข้อมูลที่ถูกดัดแปลงได้อย่างมีประสิทธิภาพมากขึ้น โดยเฉพาะในกรณีที่ภาพหรือวิดีโอมีรายละเอียดที่ซับซ้อน หรือการบิดเบือนที่ยากต่อการตรวจจับด้วยเทคนิคแบบเดิม

ดังนั้นในหัวข้อต่อไปจะกล่าวถึง โครงข่ายประสาทเทียมที่ใช้เทคนิคทรานฟอร์เมอร์ ซึ่งเป็นแนวทางที่ทันสมัยในการพัฒนาแบบจำลองการเรียนรู้เชิงลึกโดยจะอธิบายถึงหลักการทำงานและโครงสร้างของการใช้สถาปัตยกรรมทรานฟอร์เมอร์ ในบริบทของการประมวลผลภาพและวิดีโอ เพื่อให้เห็นถึงแนวทางการประยุกต์ใช้เทคโนโลยีนี้ในงานด้านต่าง ๆ อย่างเป็นระบบและมีประสิทธิภาพ

2.3 โครงข่ายประสาทเทียมที่ใช้เทคนิคทรานฟอร์เมอร์ (Transformer)

โครงข่ายประสาทเทียมที่ใช้เทคนิคทรานฟอร์เมอร์ (Transformer) คือโครงสร้างเครือข่ายประสาทเทียมที่ได้รับการออกแบบมาเพื่อจัดการกับข้อมูลลำดับ (Sequence Data) โดยมีเป้าหมายเพื่อแก้ไขข้อจำกัดของโครงสร้างแบบดั้งเดิม เช่น Recurrent Neural Networks (RNNs) และ Convolutional Neural Networks (CNNs) โดยเฉพาะอย่างยิ่งในด้านความสามารถในการประมวลผลข้อมูลลำดับที่ยาวและการคำนวณแบบขนาน ซึ่งเปิดตัวครั้งแรกในงานวิจัยชื่อว่า "Attention is All You Need" โดย Vaswani และคณะ ในปี 2017 (21) ซึ่ง Transformer ได้รับการพัฒนาให้ทำงานได้อย่างมีประสิทธิภาพทั้งในงานประมวลผลภาษา (Natural Language Processing - NLP) และการสร้างแบบจำลองลำดับในรูปแบบอื่น ๆ เช่น การแปลภาษา การสร้างข้อความ และการวิเคราะห์ภาพ

2.3.1 สถาปัตยกรรมโมเดล (Model Architecture)

Transformer ถูกออกแบบขึ้นมาเพื่อใช้ในงาน Sequence Transduction เช่น การแปลภาษาแบบประสาทเทียม (Neural Machine Translation) โดยมีความสามารถเด่นในการแปลงลำดับข้อมูลนำเข้า (Input Sequence) ให้เป็นลำดับข้อมูลส่งออก (Output Sequence) ได้อย่างมีประสิทธิภาพ และถือเป็นโมเดลลำดับตัวแรกที่อาศัย Self-Attention อย่างสมบูรณ์ในการประมวลผลข้อมูล แทนที่จะใช้ RNNs หรือ Convolutions แบบดั้งเดิม Transformer ยังคงรักษาโครงสร้างแบบ Encoder-Decoder Model ซึ่งประกอบด้วยสองส่วนหลัก

- **Encoder** ทำหน้าที่รับข้อมูลนำเข้า (Input) และแปลงข้อมูลนั้นเป็นการแทนค่าทางคณิตศาสตร์ในรูปแบบเมทริกซ์ (Encoded Representation)

- **Decoder** รับการแทนค่าจาก Encoder และสร้างผลลัพธ์ออกมาแบบลำดับ (Iteratively)

ทั้งส่วน Encoder และ Decoder ต่างประกอบด้วยหลายชั้นซ้อนกัน (ในโครงสร้างดั้งเดิมมี 6 ชั้น) และ แต่ละชั้นมีโครงสร้างเหมือนกันทั้งหมด ข้อมูลจะถูกส่งผ่าน Encoder แต่ละชั้นแบบต่อเนื่อง และผลลัพธ์จาก Encoder สุดท้ายจะถูกส่งต่อไปยัง Decoder ทุกชั้น เพื่อสร้างลำดับผลลัพธ์ (Output Sequence)

2.3.1.1 Attention Module

ในงาน Sequence-to-Sequence Learning เช่น การแปลภาษา แบบจำลอง seq2seq (22)แบบดั้งเดิมที่ใช้ RNN มักเผชิญปัญหาคอขวด (Bottleneck) เนื่องจากข้อมูลบางส่วนสูญหายระหว่างการประมวลผลข้อมูลลำดับยาว ๆ เทคนิค Attention ถูกนำมาใช้เพื่อ

แก้ปัญหานี้ โดยช่วยให้ Decoder สามารถมุ่งความสนใจไปยังส่วนของข้อมูลนำเข้าที่สำคัญได้โดยตรง ลดปัญหา Vanishing Gradient และช่วยให้โมเดลจับความสัมพันธ์ในข้อมูลได้แม่นยำมากขึ้น

อย่างไรก็ตาม Attention Module ใน Transformer มีความแตกต่างจาก Attention แบบเดิมใน seq2seq เนื่องจาก Transformer ใช้ Self-Attention ซึ่งช่วยให้ทุกหน่วยข้อมูลในลำดับสามารถโฟกัสไปยังหน่วยข้อมูลอื่น ๆ ภายในลำดับเดียวกัน โดยทำให้โมเดลเข้าใจความสัมพันธ์ระหว่างคำหรือหน่วยข้อมูลทุกคำได้อย่างละเอียดและทั่วถึง โดยการทำ Scaled Dot-Product Attention และเสริมด้วย Multi-Head Attention ซึ่งช่วยให้โมเดลสามารถจับความสัมพันธ์ได้หลากหลายมุมมองและมีประสิทธิภาพมากขึ้น

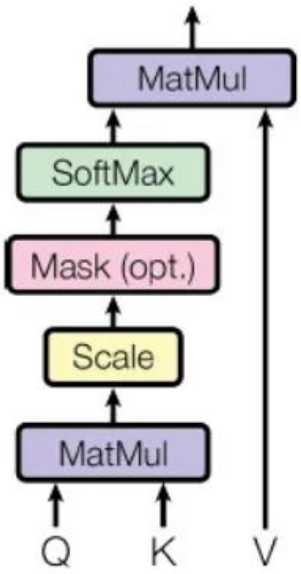
1) Scaled Dot-Product Attention วิธีการคำนวณความสัมพันธ์ใน Self-Attention ใช้กระบวนการ Scaled Dot-Product Attention ซึ่งทำงานดังนี้ คำนวณความสัมพันธ์ระหว่าง Query (Q) และ Key (K) โดยใช้การคูณเมทริกซ์ (Dot Product) แบ่งผลลัพธ์ด้วยค่ารากที่สองของมิติ $\sqrt{d_k}$ เพื่อลดค่าขนาดใหญ่ผ่านฟังก์ชัน Softmax เพื่อแปลงค่าเป็นการกระจายความน่าจะเป็นนำค่าความน่าจะเป็นไปคูณกับ Value (V) เพื่อสร้างผลลัพธ์ที่เน้นข้อมูลสำคัญ

สมการของ Scaled Dot-Product Attention

$$\mathrm{Attention}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \mathrm{softmax}\left(\frac{\mathrm{QK^T}}{\sqrt{\mathrm{d_k}}}\right)\mathrm{V} \quad (1)$$

โดยที่

- Q คือ เมทริกซ์ที่ประกอบด้วยเวกเตอร์ Query
- K คือ เมทริกซ์ที่ประกอบด้วยเวกเตอร์ Key
- V คือ เมทริกซ์ที่ประกอบด้วยเวกเตอร์ Value
- d_k คือ มิติของเวกเตอร์ Key (หรือ Query)



ภาพประกอบ 3 : Scaled Dot-Product Attention

ที่มา:(21)

2) Multi-Head Attention เป็นกระบวนการที่ใช้หลายชุดของ Scaled Dot-Product Attention โดยแบ่งการคำนวณเป็นหลาย "หัว" (Heads) แต่ละหัวจะทำการคำนวณ Query (Q), Key (K) และ Value (V) ในมิติย่อยที่แตกต่างกัน ช่วยให้โมเดลสามารถเรียนรู้และดึงข้อมูลสำคัญจากหลายมุมมองของข้อมูลลำดับ (Sequence Data) พร้อมกันการใช้หลายหัวทำให้ Multi-Head Attention สามารถจับความสัมพันธ์ในเชิงลึกและครอบคลุมมากขึ้น เมื่อเทียบกับการใช้ Attention เพียงหัวเดียว

ขั้นตอนการทำงานของ Multi-Head Attention

1. การแปลง Query, Key, และ Value แทนที่จะใช้เพียงเวกเตอร์เดียวสำหรับ Q , K , V การคำนวณใน Multi-Head Attention จะทำการแปลง Q , K , V ให้เป็นหลายชุดของเวกเตอร์ย่อยที่มีขนาดเล็กลง โดยใช้เมทริกซ์การแปลง W_i^Q, W_i^K, W_i^V แต่ละชุดเวกเตอร์ย่อยจะถูกใช้ในการคำนวณ Attention เฉพาะตัว เพื่อดึงลักษณะเฉพาะของข้อมูลที่แตกต่างกัน

2. การคำนวณ Attention หลายหัว (Parallel Attention) หลังจากการแปลง Q , K , V แล้ว จะมีการคำนวณ Attention แยกสำหรับแต่ละหัว โดยกระบวนการคำนวณในแต่ละหัวจะเหมือนกับ Scaled Dot-Product Attention แต่ละหัวจะทำงานในมิติย่อยของข้อมูลที่แตกต่างกัน ช่วยให้สามารถเรียนรู้ความสัมพันธ์หลายมิติได้พร้อมกัน

3. การรวมผลลัพธ์จากทุกหัว (Concatenation) เมื่อได้ผลลัพธ์จากทุกหัวแล้ว $(\text{head}_1, \text{head}_2, \dots, \text{head}_h)$ จะนำมารวมกันผ่านการเชื่อมต่อ (Concatenation) ผลลัพธ์ที่ได้จะถูกแปลงอีกครั้งผ่าน เมทริกซ์การแปลงรวม W^O เพื่อสร้างผลลัพธ์สุดท้ายที่พร้อมสำหรับการประมวลผลในขั้นตอนถัดไปของโมเดล

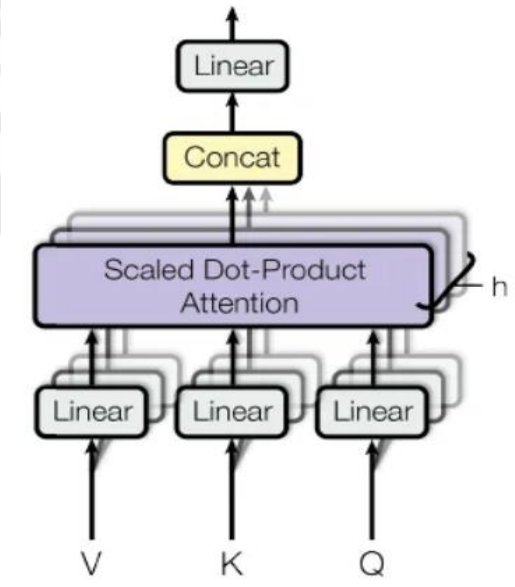
สมการของ Multi-Head Attention

$$\text{MutiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_h)W^O \quad (2)$$

โดยที่

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (3)$$

- W_i^Q, W_i^K, W_i^V คือ เมทริกซ์การแปลงที่ใช้สำหรับ Query, Key, และ Value สำหรับแต่ละหัว
- W^Oคือ เมทริกซ์การแปลงที่ใช้หลังจากการรวมผลลัพธ์ของทุกๆ หัว
-

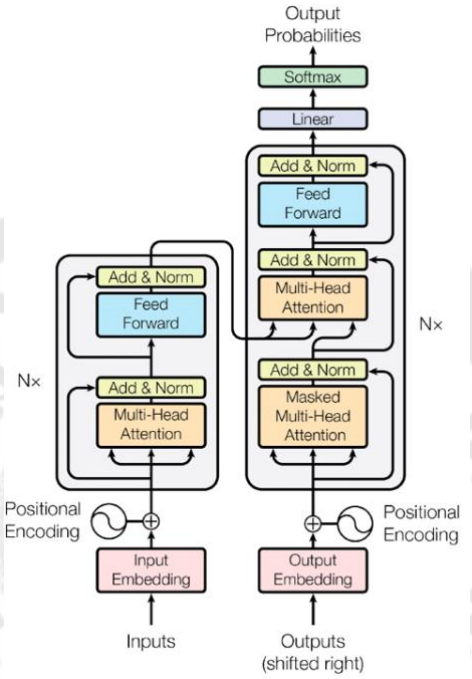


ภาพประกอบ 4 : Multi-Head Attention

ที่มา:(21)

2.1.1.2 Transformer model

แบบจำลอง Transformer นี้ยังคงมีโครงสร้างที่แบ่งเป็นสองฝั่ง โดยฝั่งซ้ายจะเป็น encoder ซึ่งรับ input sequence เข้ามา ส่วนฝั่งขวาคือ decoder ที่รับ output sequence ในขั้นตอนการฝึกอบรม output sequence จะถูกเลื่อนไปทางขวาหนึ่งตำแหน่ง ซึ่งเป็นกระบวนการฝึกแบบ teacher forcing



ภาพประกอบ 5 : Transformer model

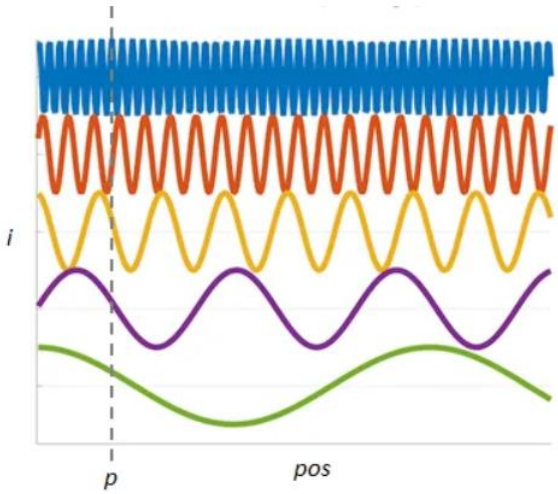
ที่มา:(21)

1) Embedding & Positional Encoding ขั้นตอนถัดไปคือ การทำ embedding ซึ่งเป็นขั้นตอนพื้นฐานสำหรับงาน NLP แทบทุกประเภท โดยสามารถใช้วิธีการทำ embedding ได้หลายรูปแบบ ในงานวิจัยนี้ใช้วิธี Byte-Pair Encoding (23) ซึ่งเป็นวิธีที่ได้รับความนิยมในงาน neural machine translation ในปัจจุบัน เนื่องจากในที่นี้เรากำลังใช้เพียง Attention โดยไม่ได้ใช้ RNN หรือ CNN จึงทำให้ข้อมูลเกี่ยวกับตำแหน่งสูญหายไป จึงจำเป็นต้องใส่ Positional Encoding (24) เข้าไป เพื่อให้โมเดลสามารถรับรู้ตำแหน่งของแต่ละโทเค็นในลำดับ ซึ่งในที่นี้ใช้ periodic function เช่น sin และ cos ที่มีความถี่ต่างๆ กันในการคำนวณค่าในแต่ละตำแหน่ง ดังสมการนี้

$$\mathrm{PE}(\mathrm{pos}, 2\mathrm{i}) = \sin\left(\frac{\mathrm{pos}}{10000^{\frac{2\mathrm{i}}{\mathrm{d_{model}}}}}\right) \quad (4)$$

$$\mathrm{PE}(\mathrm{pos}, 2\mathrm{i}) = \cos\left(\frac{\mathrm{pos}}{10000^{\frac{2\mathrm{i}}{\mathrm{d_{model}}}}}\right) \quad (5)$$

โดยที่ pos คือ ตำแหน่ง และ $2i$ กับ $2i+1$ คือ มิติของ position vector ในแต่ละตำแหน่ง ซึ่งสามารถแสดงเป็นรูปแบบคร่าวๆ เพื่อให้เข้าใจได้ง่ายขึ้นตามภาพประกอบ 6 ดังนี้

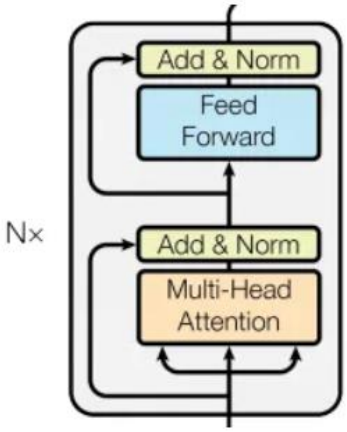


ภาพประกอบ 6 : ตัวอย่าง Positional Encoding

ที่มา: https://medium.com/mena-ai/88-3-15dcfa97763a

ในที่นี้ position vector ของตำแหน่ง p ที่ตัวที่ j จะได้จากค่าของ periodic function ที่ตำแหน่ง p โดยที่ความถี่สูงจะช่วยแยกแยะตำแหน่งที่อยู่ใกล้กัน และความถี่ต่ำจะช่วยแยกแยะตำแหน่งที่อยู่ห่างกัน การทำ Positional Encoding แบบนี้เป็นการใช้ข้อมูลเชิงความถี่มาผสมผสานกันเพื่อเพิ่มความสามารถในการรับรู้ตำแหน่งของแต่ละโทเค็นในลำดับ

2) Encoder ในแต่ละ block ของ Encoder ใน Transformer จะประกอบด้วยสามส่วนหลักๆ ที่ทำงานร่วมกันเพื่อประมวลผลข้อมูลลำดับดังนี้



ภาพประกอบ 7 : ส่วนของ Encoder

ที่มา:(21)

1. Multi-Head Attention

1.1 ส่วนแรกของแต่ละ block คือ Multi-Head Attention ซึ่งเป็น sub-layer หลัก ที่ช่วยให้โมเดลสามารถโฟกัสไปที่ข้อมูลสำคัญจากทุกตำแหน่งในลำดับข้อมูล ทั้งจากตำแหน่งก่อนหน้าและตำแหน่งอื่นๆ

1.2 ข้อมูลที่เข้าไปใน Multi-Head Attention จะประกอบด้วยสามส่วนหลัก คือ Query (Q), Key (K), และ Value (V) โดยจะคำนวณ Attention ระหว่าง Q, K, และ V เพื่อดึงข้อมูลที่สำคัญออกมา

1.3 การคำนวณนี้จะทำผ่าน หลายหัว (Multi-Head) เพื่อให้สามารถจับความสัมพันธ์จากหลายมุมมองของข้อมูลในลำดับ

2. Position-wise Feed-Forward Network

2.1 หลังจากผ่านการคำนวณใน Multi-Head Attention ผลลัพธ์จะถูกส่งไปที่ Position-wise Fully Connected Feed-Forward Network ซึ่งประกอบด้วยสองชั้น

2.2 Position-wise หมายถึงการประมวลผลข้อมูลในแต่ละตำแหน่งแยกจากกัน โดยไม่มีการพึ่งพากันระหว่างตำแหน่งข้อมูล เช่นเดียวกับ 1x1 Convolutional Network

2.3 ข้อมูลที่ผ่านการประมวลผลในเครือข่ายนี้จะถูกส่งออกไปเป็นผลลัพธ์ที่ผ่านการคำนวณแล้ว

3. Residual Connection และ Layer Normalization (25)

ทั้งใน Multi-Head Attention และ Position-wise Feed-Forward Network จะมีการใช้ Residual Connection ซึ่งหมายถึง การนำผลลัพธ์จากการคำนวณแต่ละส่วนมาบวกเข้ากับข้อมูลขาเข้าของ layer นั้นๆ เพื่อให้ข้อมูลผ่านการประมวลผลโดยไม่สูญเสียข้อมูลดั้งเดิม สูตรสำหรับการคำนวณจะเป็นดังนี้

$$\text{Output} = \text{LayerNorm}(x + \text{SubLayer}(x)) \quad (6)$$

โดยที่

- x คือ ข้อมูลขาเข้าของ layer
- SubLayer(x) คือ ผลลัพธ์ที่ได้จากการประมวลผลในแต่ละ sub-layer เช่น Multi-Head Attention หรือ Feed-Forward Network
- LayerNorm คือ กระบวนการปรับสเกลข้อมูลเพื่อให้ได้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานที่เหมาะสม

3) Decoder ในแบบจำลอง Transformer เป็นส่วนที่รับผิดชอบในการสร้างผลลัพธ์สุดท้ายจากข้อมูลที่ได้รับจาก encoder โดยใช้กระบวนการที่เรียกว่า Self-Attention และ Encoder-Decoder Attention เพื่อทำความเข้าใจข้อมูลที่ผ่านการประมวลผลจาก encoder ก่อนหน้านั้นและสร้างผลลัพธ์ที่ต้องการ การทำงานของ decoder มีลักษณะดังนี้

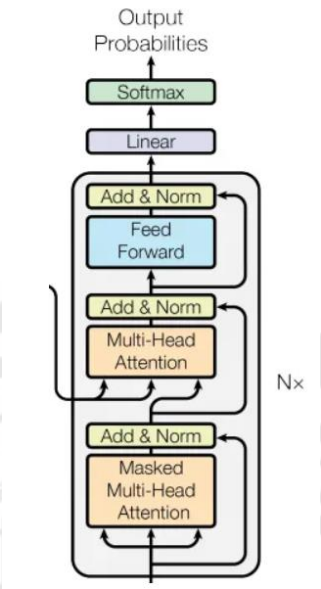
1. Masked Self-Attention: ในแต่ละ block ของ decoder มีการใช้ Masked Self-Attention เพื่อให้โมเดลสามารถคำนึงถึงคำที่ถูกประมวลผลก่อนหน้านี้ (ไม่เห็นคำในอนาคต) ซึ่งช่วยในการสร้างผลลัพธ์ทีละคำ โดยไม่สามารถเข้าถึงคำในอนาคตได้ในระหว่างการสร้าง output

2. Encoder-Decoder Attention: Layer นี้จะใช้ข้อมูลที่ได้รับจาก encoder (โดยใช้ค่าจาก Key และ Value) ร่วมกับข้อมูลที่อยู่ใน decoder (Query) เพื่อช่วยให้ decoder เข้าใจบริบทของข้อมูลจาก encoder และสร้างผลลัพธ์ที่สอดคล้องกับข้อมูลนั้น

3. Feedforward Neural Network: หลังจากผ่านการประมวลผลด้วย Attention, ข้อมูลจะผ่านขั้นตอนของ Neural Network ที่ช่วยเพิ่มความซับซ้อนในการคำนวณและตัดสินใจ

ในส่วนของ decoder จะประกอบด้วย 6 block ที่เรียงกันเทียบเท่ากับจำนวนของ encoder โดยผลลัพธ์จาก block สุดท้ายของ encoder จะถูกส่งต่อไปยัง layer ตรงกลางของแต่ละ block ใน decoder เพื่อให้ข้อมูลจาก encoder สามารถนำมาประมวลผลร่วมกับข้อมูลใน

decoder ได้อย่างมีประสิทธิภาพ ผลลัพธ์จาก block สุดท้ายของ decoder จะถูกส่งผ่านกระบวนการ softmax เพื่อแปลงข้อมูลเป็นผลลัพธ์ output ต่อไป



ภาพประกอบ 8 : ส่วนของ Decoder

ที่มา:(21)

2.3.2 ประเภทของแบบจำลอง Transformer

แบบจำลอง Transformer มีหลากหลายประเภทที่ได้รับการพัฒนาขึ้นเพื่อตอบสนองความต้องการในด้านต่างๆ ของงานประมวลผลข้อมูลและการเรียนรู้เชิงลึก (Deep Learning) โดยเฉพาะในงาน Natural Language Processing (NLP) และ Computer Vision ที่มีการใช้ประโยชน์จาก Transformer อย่างกว้างขวาง โดยแบ่งเป็นหลายประเภท ยกตัวอย่างดังนี้

2.3.2.1 Vision Transformer (ViT)

Vision Transformer (ViT) นำสถาปัตยกรรม Transformer มาเปลี่ยนไปใช้สำหรับงานการจำแนกภาพ (image classification) แทนที่จะประมวลผลภาพเป็นตารางของพิกเซล โมเดลนี้จะดูข้อมูลภาพเป็นลำดับของ patches ขนาดคงที่ ซึ่งคล้ายกับวิธีการปฏิบัติต่อคำในประโยค แต่ละ patch จะถูกปรับให้เป็นแนวนอนและฝังแบบเชิงเส้น แล้วประมวลผลตามลำดับโดยตัวเข้ารหัสของ Transformer มาตรฐาน มีการเพิ่มการฝังตามตำแหน่งเพื่อรักษาข้อมูลเชิง

พื้นที่ การใช้เทคนิค Self-attention ทั่วโลกนี้จะช่วยให้โมเดลสามารถบันทึกความสัมพันธ์ระหว่าง patch คู่ใดก็ได้โดยไม่ต้องคำนึงถึงตำแหน่ง (26)

การทำงานหลักของ ViT

1) **การแบ่งภาพเป็น Patch** ภาพจะถูกแบ่งออกเป็นหลายๆ patch ขนาดเล็กๆ (เช่น 16x16 pixels) โดยแต่ละ patch จะถูกแปลงเป็นเวกเตอร์ 1D ด้วยการ Flatten จากนั้นจึงแปลงเวกเตอร์เหล่านี้เป็นลำดับของ token (เหมือนกับคำใน NLP)

2) **Embedding Layer Patch** แต่ละอันจะถูกแปลงเป็นเวกเตอร์ที่มีขนาดเดียวกันผ่าน embedding และตำแหน่งของ patch จะถูกเข้ารหัสด้วย Positional Encoding เพื่อให้โมเดลสามารถรับรู้ลำดับตำแหน่งของ patch ภายในภาพได้ (เหมือนกับ positional encoding ใน NLP)

3) **Transformer Encoder** ลำดับของ patch embeddings ถูกส่งผ่านโมเดล Transformer Encoder ซึ่งประกอบไปด้วย layers ของ Self-Attention และ Feedforward Networks การประมวลผลนี้ช่วยให้โมเดลเรียนรู้บริบทและความสัมพันธ์ระหว่าง patch ต่างๆ ในภาพ

4) **การจำแนกประเภท** หลังจากที่ผ่านการประมวลผลโดย Transformer, output จะถูกนำไปผ่าน MLP Head (Multi-Layer Perceptron) เพื่อทำนายคลาสของภาพ

2.3.2.2 DeiT (Data-efficient Image Transformer)

DeiT เป็นการพัฒนาเพิ่มเติมจาก Vision Transformer (ViT) โดยมุ่งเน้นที่การฝึกแบบจำลอง Transformer ในงาน การจำแนกประเภทภาพ ด้วยข้อมูลที่จำกัดและมีประสิทธิภาพสูงขึ้นในด้าน Data Efficiency

การทำงานหลักของ DeiT

1) **Teacher-Student Training** ในการฝึก DeiT จะใช้วิธี teacher-student ในการเรียนรู้ โดยที่ teacher model (เช่น CNN หรือ ViT ที่ฝึกมาก่อน) จะช่วยในการสร้างข้อมูล pseudo labels สำหรับ student model (DeiT) ซึ่งเป็นโมเดลที่เราต้องการฝึก

2) **การใช้ข้อมูลเพิ่มเติม** โมเดลใช้ข้อมูลเสริม เช่น augmentation techniques เพื่อช่วยให้โมเดลเรียนรู้ได้จากข้อมูลน้อยลง โดยไม่จำเป็นต้องพึ่งพาข้อมูลจำนวนมาก

3) **Efficient Training** ซึ่ง DeiT ถูกออกแบบมาให้ใช้ทรัพยากรในการฝึกน้อยลง แต่ยังคงสามารถสร้างผลลัพธ์ที่ดีในงานจำแนกประเภทภาพ

3.2.3 Swin Transformer

Swin Transformer (Shifted Window Transformer) เป็นการปรับปรุง Vision Transformer (ViT) โดยใช้แนวคิดของ Windows ในการแบ่งภาพเพื่อเพิ่มประสิทธิภาพในงานที่มีขนาดข้อมูลใหญ่และซับซ้อน เช่น การจำแนกประเภทภาพและการตรวจจับวัตถุ

การทำงานหลักของ Swin Transformer

1) Shifting Windows แทนที่จะใช้ global self-attention ซึ่งคำนวณความสัมพันธ์ระหว่างทุกๆ patch ในภาพ (ซึ่งใช้เวลาและหน่วยความจำมาก) Swin ใช้ local self-attention ใน windows ขนาดเล็กที่ครอบคลุมกลุ่ม patch บางตัว โดยการคำนวณจะทำเฉพาะในหน้าต่าง (window) ที่ถูกเลือก

2) Shifted Windows ในแต่ละ layer ของโมเดล, windows จะถูก shifted ไปเล็กน้อย ซึ่งช่วยให้โมเดลสามารถเรียนรู้ความสัมพันธ์ระหว่าง windows ที่ไม่อยู่ในตำแหน่งเดียวกัน และช่วยให้การเรียนรู้ลักษณะของภาพมีประสิทธิภาพมากขึ้น

3) Hierarchical Structure Swin ใช้โครงสร้างที่มีการลดขนาดของ windows ในแต่ละ layer เพื่อสร้างลำดับชั้นที่มีความละเอียดต่ำลง ซึ่งช่วยให้โมเดลสามารถจัดการกับข้อมูลที่มีความซับซ้อนได้ดีขึ้น

4) Patch Partitioning เช่นเดียวกับ ViT, Swin จะเริ่มต้นโดยการแบ่งภาพเป็น patch แต่จะใช้ patch merging ในแต่ละ layer เพื่อลดความละเอียดและทำให้มีประสิทธิภาพในการประมวลผล

2.4 งานวิจัยที่เกี่ยวข้อง

ปี 2021 งานวิจัยของ Dosovitskiy และ คณะ (27) ได้นำเสนอแบบจำลอง Vision Transformer (ViT) ซึ่งเป็นการประยุกต์ใช้สถาปัตยกรรม Transformer จากงานประมวลผลภาษา (NLP) มาสู่การจำแนกภาพ โดยมีเป้าหมายเพื่อแสดงให้เห็นว่าแบบจำลอง Transformer สามารถทำงานได้ดีในงานจำแนกภาพโดยไม่ต้องพึ่งพาโครงสร้างเฉพาะสำหรับภาพ (Inductive Bias) ที่มีในแบบจำลอง CNNs งานวิจัยนี้ใช้การแบ่งภาพออกเป็นแพตช์ขนาดเล็ก 16x16 พิกเซล แล้วแปลงแต่ละแพตช์เป็นเวกเตอร์เชิงเส้น จากนั้นป้อนเข้าสู่ Transformer Encoder ซึ่งเป็นโครงสร้างเดียวกับที่ใช้ใน NLP พร้อมกับการใช้ Position Embedding เพื่อรักษาความสัมพันธ์เชิงพื้นที่ในภาพ ผลการทดลองแสดงให้เห็นว่าแบบจำลอง ViT มีประสิทธิภาพสูงในงานจำแนกภาพ โดยเฉพาะอย่างยิ่งเมื่อฝึกฝนบนชุดข้อมูลขนาดใหญ่ เช่น JFT-300M (303 ล้านภาพ) และ ImageNet-21k (14 ล้านภาพ) ซึ่งช่วยให้แบบจำลองบรรลุความแม่นยำสูงถึง 88.55% บน ImageNet, 94.55% บน CIFAR-100, และ 99.68% บน Oxford Flowers-102 นอกจากนี้ แบบจำลอง ViT ยังแสดงให้เห็นว่ามีศักยภาพเหนือกว่าแบบจำลอง CNNs ชั้นนำ เช่น ResNet และ EfficientNet ทั้งในด้านความแม่นยำและการใช้ทรัพยากรการคำนวณที่น้อยกว่า งานวิจัยนี้ยังเน้นถึงความสำคัญของการใช้ข้อมูลขนาดใหญ่ในการฝึก ViT เพื่อชดเชยการขาด Inductive Bias ที่มีอยู่ใน CNNs และแสดงให้เห็นว่า ViT สามารถถ่ายโอนความรู้ไปยังชุดข้อมูลขนาดเล็กได้อย่างมีประสิทธิภาพ นอกจากนี้ยังมีข้อเสนอแนะในการขยายการใช้งาน ViT ไปสู่ด้านอื่น ๆ เช่น Object Detection, Segmentation, และการฝึกแบบ Self-Supervised Learning เพื่อลดการพึ่งพาข้อมูลที่มีการติดป้ายกำกับ งานวิจัยนี้จึงถือเป็นหลักฐานสำคัญที่ยืนยันว่า Transformers สามารถนำมาประยุกต์ใช้ในงานวิเคราะห์ภาพได้อย่างมีประสิทธิภาพ และยังเป็นพื้นฐานสำหรับการพัฒนาแบบจำลองใหม่ ๆ ในอนาคตอีกด้วย

ปี 2021 งานวิจัยของ Touvron และ คณะ (28) ได้นำเสนอแนวทางใหม่ในการพัฒนา Vision Transformers (ViT) สำหรับงานการจำแนกภาพ โดยมุ่งลดข้อจำกัดด้านการใช้ข้อมูลขนาดใหญ่และทรัพยากรการคำนวณจำนวนมาก ซึ่งเป็นอุปสรรคสำคัญของแบบจำลองประเภทนี้ ซึ่งงานวิจัยนี้ได้นำเสนอแบบจำลอง Data-Efficient Image Transformers (DeiT) ที่สามารถฝึกฝนได้โดยใช้ชุดข้อมูล ImageNet เพียงอย่างเดียว โดยใช้ทรัพยากรเพียง 8 GPUs และใช้เวลาฝึกฝนน้อยกว่า 3 วัน ผลลัพธ์ที่ได้มีความแม่นยำสูงถึง 85.2% (Top-1 Accuracy) บน ImageNet ซึ่งสามารถแข่งขันหรือเหนือกว่าแบบจำลอง ConvNets ชั้นนำ เช่น ResNet และ EfficientNet จุดเด่นสำคัญของงานวิจัยนี้คือการพัฒนา Distillation Token ซึ่งเป็นกลยุทธ์การถ่ายโอนความรู้ (Knowledge Distillation) แบบเฉพาะสำหรับ Transformers โดยการเพิ่ม Distillation Token เข้ามาในแบบจำลอง ทำให้ Transformer สามารถเรียนรู้จากผลลัพธ์ของ แบบจำลองครู (Teacher Model) ได้อย่างมีประสิทธิภาพมากขึ้น โดยเฉพาะเมื่อใช้ CNNs เป็นแบบจำลองครู ผลการทดลองแสดงให้เห็นว่าเทคนิคนี้ให้ประสิทธิภาพเหนือกว่าวิธีการถ่ายโอนความรู้แบบ Soft Distillation แบบเดิมอย่างชัดเจน อีกทั้งยังช่วยให้แบบจำลองสามารถถ่ายโอนความรู้ไปยังงานอื่น ๆ ได้อย่างมีประสิทธิภาพ เช่น การจำแนกภาพในชุดข้อมูล CIFAR-10, CIFAR-100, และ Stanford Cars นอกจากนี้งานวิจัยยังเน้นการใช้เทคนิคที่ช่วยเพิ่มประสิทธิภาพการฝึก เช่น การปรับข้อมูล (Data Augmentation) เช่น RandAugment, Mixup, และ CutMix รวมถึงการใช้ตัวปรับพารามิเตอร์การเรียนรู้ (Optimizer) อย่าง AdamW และเทคนิค Stochastic Depth เพื่อเพิ่มความเสถียรในการเรียนรู้ งานวิจัยนี้จึงแสดงให้เห็นว่าแบบจำลอง DiT สามารถก้าวข้ามข้อจำกัดด้านข้อมูลและทรัพยากรการคำนวณ และมีศักยภาพสูงที่จะกลายเป็นทางเลือกสำคัญในงานด้านการจำแนกภาพในอนาคต เนื่องจากมีทั้งประสิทธิภาพสูงและความสามารถทั่วไปของแบบจำลองที่ดี

ปี 2021 งานวิจัยของ Liu และ คณะ (29) ได้นำเสนอแบบจำลอง Swin Transformer เพื่อปรับปรุงประสิทธิจาก Vision Transformers โดยออกแบบให้เป็นแบ็กโบนอเนกประสงค์สำหรับงาน Computer Vision โดยมุ่งเน้นการแก้ปัญหาความแตกต่างของข้อมูลภาพและข้อความ เช่น ความละเอียดที่สูงของพิกเซลในภาพ และความหลากหลายของขนาดวัตถุในงานวิชั่น แบบจำลอง Swin Transformer ใช้โครงสร้างแบบลำดับชั้น (Hierarchical Architecture) ซึ่งสร้างแผนที่คุณลักษณะ (feature maps) แบบลำดับชั้น โดยเริ่มจากการแบ่งภาพเป็นส่วนย่อย (patches) และรวมข้อมูลในเลเยอร์ที่ลึกขึ้น เพื่อรองรับการทำนายที่หลากหลายสเกล นอกจากนี้ยังมีการประยุกต์ใช้เทคนิค Shifted Windows ที่ช่วยคำนวณ self-attention เฉพาะในหน้าต่างที่ไม่ทับซ้อนกัน และเพิ่มความสามารถในการเชื่อมต่อระหว่างหน้าต่างด้วยการเลื่อน ช่วยลดความซับซ้อนในการคำนวณและลดเวลาแฝง (Latency) โมเดลนี้มีความซับซ้อนเชิงเส้นเมื่อเทียบกับขนาดของภาพ ทำให้เหมาะสำหรับการประมวลผลภาพที่มีความละเอียดสูง ผลการทดลองแสดงให้เห็นว่า Swin Transformer มีประสิทธิภาพที่เหนือกว่าแบบจำลองก่อนหน้าในหลากหลายด้าน เช่น ด้าน Image Classification บรรลุความแม่นยำ 87.3% บน ImageNet-1K ด้าน Object Detection ทำคะแนน 58.7 Box AP และ 51.1 Mask AP บน COCO test-dev ซึ่งเพิ่มขึ้น +2.7 และ +2.6 เมื่อเทียบกับ state-of-the-art เดิม และด้าน Semantic Segmentation ทำคะแนน 53.5 mIoU บน ADE20K val สูงขึ้น +3.2 จากผลงานก่อนหน้า ด้วยคุณสมบัติและผลลัพธ์ที่โดดเด่น Swin Transformer ได้พิสูจน์ให้เห็นถึงศักยภาพในการเป็นแบ็กโบนสำหรับงานวิชั่นที่หลากหลาย ไม่ว่าจะเป็นการพัฒนาแบบจำลองที่ผสานการประมวลผลภาพและภาษาธรรมชาติ (NLP) หรือการประมวลผลภาพที่มีความละเอียดสูงและรองรับการทำนายแบบหลายสเกล ทำให้งานวิจัยจำนวนมากเริ่มนำโครงสร้างนี้มาประยุกต์ใช้ในการตรวจจับภาพปลอม (Deepfake) และคาดว่าจะมีบทบาทสำคัญในการพัฒนาเทคโนโลยีดังกล่าวในอนาคต

ปี 2022 Nguyen และ คณะ (1) ได้นำเสนอบทความวิจัยทำการสำรวจเทคนิคในงานวิจัยต่างๆ ที่มีอยู่ในปัจจุบันที่ใช้เทคโนโลยีการเรียนรู้เชิงลึก (Deep learning) ในการสร้างและตรวจจับภาพปลอม และ วีดีโอปลอม ที่ถูกปลอมแปลงให้เหมือนจริง (Deepfake) ส่วนแรกการสร้าง deepfake จะใช้เทคโนโลยี Generative Adversarial Networks (GANs) ที่เป็นเครื่องมือหลักในการสร้างเนื้อหาปลอมแปลง GANs โดยจะประกอบด้วยสองแบบจำลองหลักคือ *Generator* ซึ่งมีหน้าที่สร้างเนื้อหาปลอม และ *Discriminator* ที่ทำหน้าที่ตรวจจับว่าเนื้อหานั้นจริงหรือปลอม โดยทั้งสองแบบจำลองจะการทำงานร่วมกันทำให้การสร้าง Deepfake ออกมามีความซับซ้อนและเหมือนจริงมากยิ่งขึ้น นอกจากนี้ยังมีการใช้เทคโนโลยี Variational Autoencoders (VAEs) ซึ่งเป็นแบบจำลองที่ใช้ในการเข้ารหัส และถอดรหัสข้อมูลเพื่อสร้างภาพที่ดูเป็นธรรมชาติ รวมถึงเทคนิคอื่นๆ เช่น StyleGAN , Autoencoders เป็นต้น ส่วนการตรวจจับ deepfake มีการพัฒนาแบบจำลองต่างๆ ที่ใช้ในการตรวจจับภาพ ส่วนใหญ่จะใช้เทคนิคหลักคือ CNNs ,SVM , VGG , LSTM , LRNU , PRNU , Scnet , ResNet-50 และอื่นๆ ซึ่งเป็นเทคนิคที่มีประสิทธิภาพสูงในการวิเคราะห์ deepfake งานวิจัยนี้ได้ทำการสำรวจเทคนิคโดยรวมในการสร้าง การตรวจจับภาพ และวีดีโอปลอม ซึ่งปัจจุบันภาพและวีดีโอปลอม มีแนวโน้มที่จะความซับซ้อนและตรวจจับได้ยากมากยิ่งขึ้น ด้วยเหตุนี้จึงจำเป็นต้องพัฒนาวิธีการตรวจจับอย่างต่อเนื่องเพื่อเพิ่มความแม่นยำและประสิทธิภาพในการตรวจจับภาพปลอมในอนาคต

ปี 2022 Sharma J และคณะ (30) ได้ดำเนินการวิจัยเกี่ยวกับการตรวจจับภาพใบหน้าปลอม (Deepfake) ซึ่งถูกสร้างขึ้นโดยอาศัยเทคนิค GANs โดยมีวัตถุประสงค์เพื่อประยุกต์ใช้แนวทางการเรียนรู้เชิงลึก (Deep Learning) ในการจำแนกภาพใบหน้าจริงและปลอม งานวิจัยดังกล่าวได้ประเมินประสิทธิภาพของโมเดลบนชุดข้อมูลมาตรฐานจำนวน 3 ชุด ได้แก่ 1) ชุดข้อมูล 140k Real and Fake Faces ประกอบด้วยภาพจำนวน 140,000 ภาพ 2) ชุดข้อมูล Real and Fake Face Detection ประกอบด้วยภาพจำนวน 2,041 ภาพ และ 3) ชุดข้อมูล Fake Faces ซึ่งประกอบด้วยภาพจำนวน 6,400 ภาพ ผู้วิจัยได้เลือกใช้สถาปัตยกรรมโมเดล ได้แก่ ResNet50, VGG16 และ CNN เพื่อเปรียบเทียบสมรรถนะของโมเดลแต่ละแบบ โดยตั้งค่าพารามิเตอร์ให้เหมือนกันในทุกโมเดล ได้แก่ จำนวนรอบการฝึก (epochs) จำนวน 20 รอบ , ขนาดชุดข้อมูลย่อย (batch size) เท่ากับ 64 ,ใช้ฟังก์ชัน activation แบบ ReLU และ optimizer แบบ Adam ผลการทดลองแสดงให้เห็นว่า ชุดข้อมูล 140k Real and Fake Faces ส่งผลให้โมเดลมีค่าความแม่นยำโดยรวมดีที่สุดเมื่อเปรียบเทียบกับชุดข้อมูลอื่น ๆ โดยโมเดล ResNet50 ให้ค่าความแม่นยำสูงสุดที่ 93.98%, โมเดล VGG16 ให้ค่าความแม่นยำที่ 86.63%, และโมเดล CNN ให้ค่าความแม่นยำที่ 53.25% ซึ่งแสดงให้เห็นว่าทั้ง ResNet50 และ VGG16 มีศักยภาพในการตรวจจับภาพปลอมได้อย่างมีประสิทธิภาพ โดยเฉพาะเมื่อใช้งานร่วมกับชุดข้อมูลที่มีขนาดใหญ่และมีความหลากหลายสูง ต่อมางานวิจัยได้ดำเนินการสร้าง โมเดลรวม (Ensemble Model) โดยผสานการทำงานของโมเดลทั้งสาม ได้แก่ ResNet50, VGG16 และ CNN เข้าด้วยกัน ผลการประเมินพบว่า โมเดลรวมที่ได้รับสามารถบรรลุค่าความแม่นยำสูงสุดที่ 98.79% บนชุดข้อมูล 140k Real and Fake Faces ซึ่งสะท้อนถึงประสิทธิภาพที่เหนือกว่าโมเดลเดี่ยวในส่วนของข้อเสนอแนะ งานวิจัยระบุว่า ในอนาคตควรมีการขยายขนาดและความหลากหลายของชุดข้อมูลที่ใช้ในการฝึกและทดสอบ เพื่อยกระดับความสามารถในการจำแนกของโมเดลให้ดียิ่งขึ้น นอกจากนี้ยังมีแผนในการต่อยอดงานวิจัยไปสู่การตรวจจับวิดีโอใบหน้าปลอม รวมถึงการประเมินประสิทธิภาพของโมเดลบนภาพที่มีความละเอียดต่ำหรือถ่ายในสภาวะแสงน้อยอีกด้วย

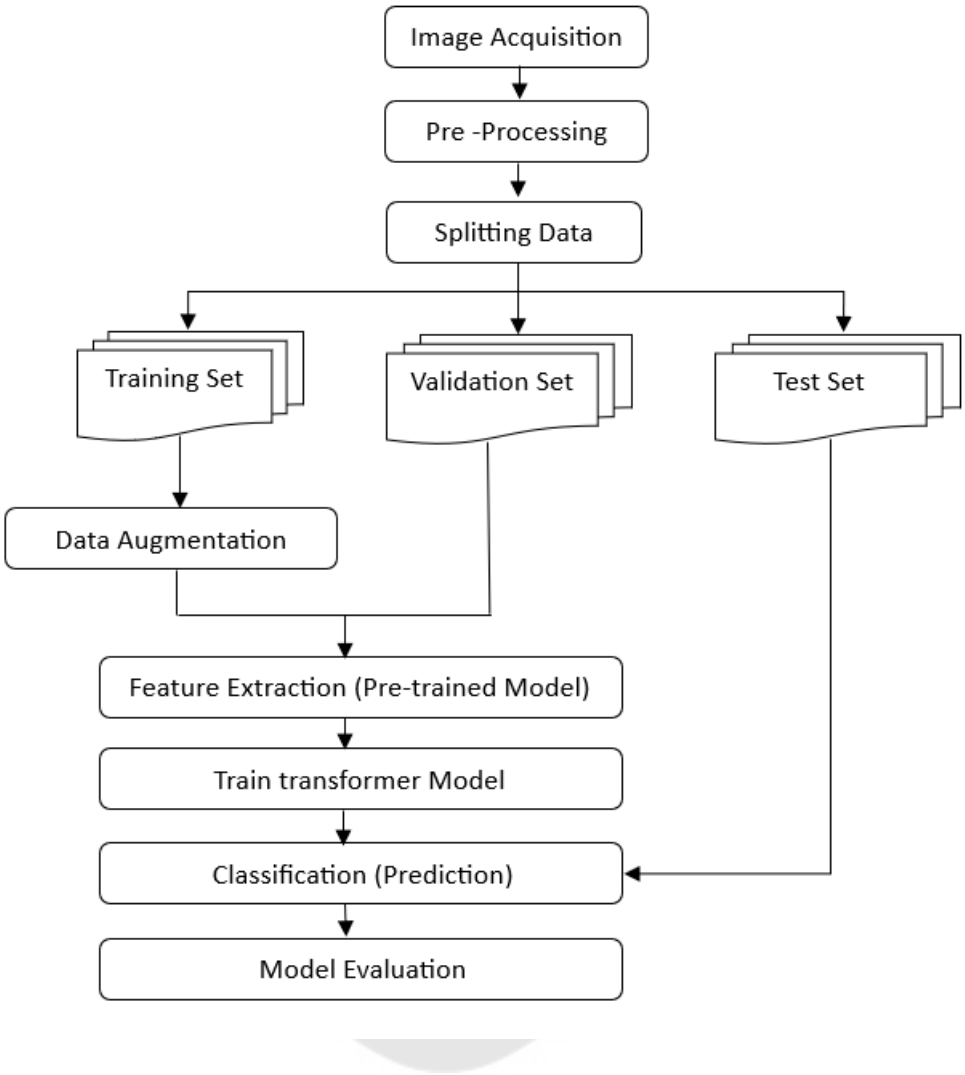
ปี 2024 Ghita และ คณะ (31) ได้นำเสนอจากบทความวิจัยพบว่าปัจจุบันแนวโน้ม deepfake มีการสร้างภาพ และ วีดีโอที่ถูกปลอมแปลงได้แนบเนียนและเหมือนภาพจริงมากยิ่งขึ้น ซึ่งเทคโนโลยีของ deepfake มีการพัฒนาอย่างต่อเนื่อง โดยวิธีที่ใช้ในการตรวจจับ deepfake ส่วนใหญ่จะใช้เทคนิคแบบจำลอง CNNs ซึ่งงานวิจัยนี้จะนำเสนอวิธีการอีกรูปแบบหนึ่งในการตรวจจับ deepfake คือ Vision Transformer (ViT) นำข้อมูลภาพจาก Kaggle ที่ประกอบด้วยภาพจริงและภาพปลอม 190,345 ภาพ ซึ่งชุดข้อมูลถูก แบ่งออกเป็น ชุดข้อมูลฝึก (Training set) เท่ากับ 80% และ ชุดข้อมูลทดสอบ (Test set) เท่ากับ 20% โดยตั้งค่าพารามิเตอร์ learning rate ที่ 0.001, weight decay ที่ 0.0001 และ batch size ที่ 256 จากการทดสอบเบื้องต้น ใช้ข้อมูลย่อย 5,850 ภาพ ประกอบด้วย deepfake 2,531 ภาพ และภาพจริง 3,319 ภาพ โดยจะย่อขนาดภาพเป็น 72x72 พิกเซล และแบ่งเป็น patch ขนาด 6x6 พิกเซล เพื่อนำเข้าสู่แบบจำลอง โดยใช้ 800 epochs ได้ค่า accuracy ของ training set เข้าใกล้ 1 แต่ validation set มีค่า accuracy ประมาณ 0.75 แสดงว่าข้อมูลเกิดการ overfitting ส่วนขั้นตอนการฝึกโมเดล ใช้ชุดข้อมูล 40,000 ภาพ ประกอบด้วย deepfake 20,000 ภาพ และภาพจริง 20,000 ภาพ โดยใช้ epochs 200 ผลลัพธ์โดยรวมของแบบจำลอง Overall accuracy เท่ากับ 89.91% ซึ่งอยู่ในระดับเดียวกับผลลัพธ์จากงานวิจัยอื่น ๆ ที่ใช้เทคนิคการเรียนรู้เชิงลึก (deep learning) จากผลการทดสอบแสดงให้เห็นว่าขนาดของชุดข้อมูลมีผลกระทบอย่างมากต่อประสิทธิภาพของแบบจำลอง ในอนาคตงานวิจัยจะใช้ชุดข้อมูลขนาดใหญ่ขึ้นเพื่อพัฒนาแบบจำลองต่อไปซึ่งคาดว่าจะได้ผลลัพธ์ที่ดียิ่งขึ้น รวมถึงการทดสอบประสิทธิภาพของแบบจำลองโดยใช้พารามิเตอร์ในการฝึกที่แตกต่างกันออกไป เพื่อให้ผลลัพธ์ของแบบจำลองมีประสิทธิภาพดียิ่งขึ้นในอนาคต

บทที่ 3
วิธีการดำเนินการวิจัย

กระบวนการทำงานวิจัยทั้งหมดในครั้งนี้ มีจุดประสงค์เพื่อศึกษาและนำวิธีการมาปรับใช้กับชุดข้อมูลเพื่อทดลองและเปรียบเทียบประสิทธิภาพของแบบจำลองด้วยการใช้โครงข่ายประสาทเทียบแบบ Transformers ผู้วิจัยได้ดำเนินการตามขั้นตอนวิจัยตามขั้นตอนดังนี้

1. กระบวนการทำงานของแบบจำลอง
2. การเตรียมข้อมูลสำหรับวิจัย (Data Acquisition & Data Filtering)
3. การเตรียมข้อมูลก่อนเข้าแบบจำลอง (Data preprocessing)
4. การสร้างแบบจำลอง (Modeling)
5. การประเมิณผล (Evaluation)

3.1 กระบวนการทำงานของแบบจำลอง



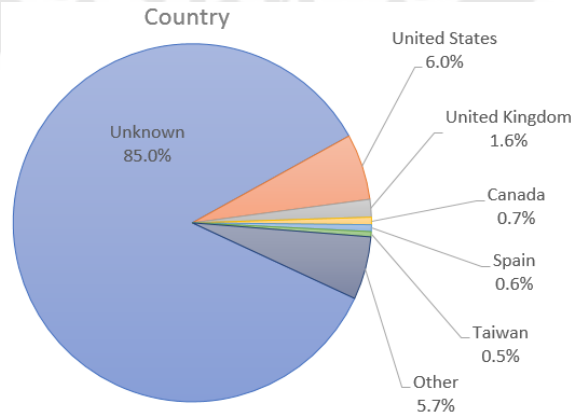
ภาพประกอบ 9 ภาพการดำเนินงานกับข้อมูลในการทำวิจัย

จากภาพประกอบ 9 งานวิจัยเริ่มจากการเก็บรวบรวมข้อมูลภาพ (Image Acquisition) และ เตรียมข้อมูลภาพ (Pre-Processing) โดยปรับขนาดและทำ Normalization ของข้อมูล จากนั้น แบ่งข้อมูล (Splitting Data) เป็น ชุดข้อมูลฝึก (Training Set), ชุดข้อมูลตรวจสอบ (Validation set) และ ชุดข้อมูลทดสอบ (Test set) สำหรับข้อมูลชุดฝึกจะมีการทำ Data Augmentation เพื่อให้ชุดข้อมูลฝึกมีความหลากหลายมากขึ้น จากนั้นเข้าสู่ขั้นตอนถัดไป นำข้อมูลชุดข้อมูลฝึก และ ชุดข้อมูลตรวจสอบ ทำ Pre-trained model ขั้นตอนถัดมาคือ การเรียนรู้แบบจำลอง (Model Training) โดยใช้เทคนิค การถ่ายโอนการเรียนรู้ (Transfer Learning) เพื่อฝึก

แบบจำลองให้เหมาะสมกับข้อมูลภาพที่เตรียมไว้ จากนั้นทำ การจำแนกประเภท (Classification) โดยใช้ชุดข้อมูลทดสอบ เพื่อตรวจสอบผลลัพธ์ สุดท้ายคือ การประเมินประสิทธิภาพ (Model Evaluation) โดยใช้ตัวชี้วัด เช่น Accuracy, Precision, Recall และ F1-Score เพื่อวัดความสามารถของแบบจำลองในการจำแนกข้อมูล

3.2 การเตรียมข้อมูลสำหรับวิจัย (Data Acquisition & Data Filtering)

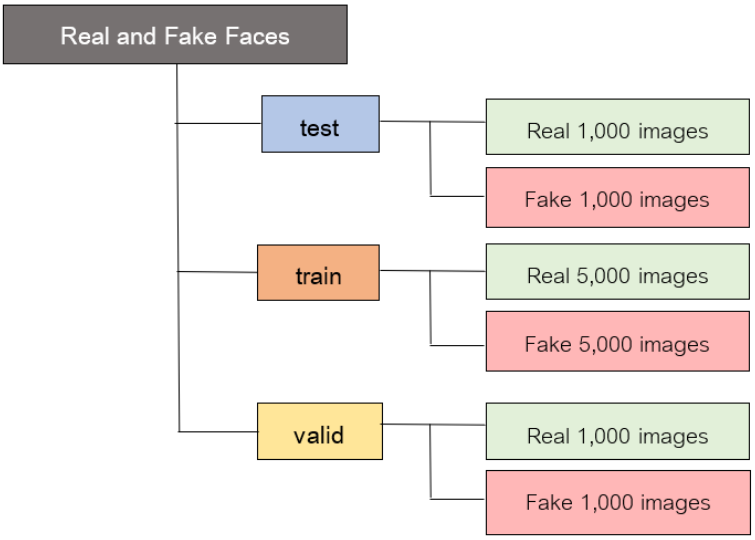
ในงานวิจัยใช้ชุดข้อมูล 140k Real and Fake Faces จาก Kaggle ชุดข้อมูลมีภาพทั้งหมดจำนวน 140,000 ภาพ ประกอบด้วยภาพใบหน้าจริงจำนวน 70,000 ภาพจากชุดข้อมูล Flickr ที่ถูกรวบรวมโดย Nvidia (32) และภาพใบหน้าปลอมจำนวน 70,000 ภาพ ที่สร้างขึ้นโดยวิธี StyleGAN ข้อมูลภาพทั้งหมดในงานวิจัยนี้ภาพชุดข้อมูลเป็นขนาดภาพเป็น 256x256 พิกเซล อยู่ในรูปไฟล์ JPG ทั้งนี้ ชุดข้อมูลมีความหลากหลายทางด้านบุคคล โดยเฉพาะในส่วนของภาพใบหน้าจริงจากชุด Flickr ซึ่งมีแหล่งที่มาจากบุคคลหลากหลายเชื้อชาติ เช่น สหรัฐอเมริกา สหราชอาณาจักร แคนาดา สเปน และไต้หวัน ดังแสดงในภาพประกอบที่ 10 แม้ว่าข้อมูลส่วนใหญ่จะไม่ได้ระบุสัญชาติของบุคคลอย่างชัดเจน แต่จากลักษณะภาพที่ปรากฏสามารถสะท้อนให้เห็นถึงความหลากหลายทางเชื้อชาติและลักษณะทางกายภาพของบุคคลได้อย่างชัดเจน จึงนับว่าเป็นชุดข้อมูลที่มีความเหมาะสมสำหรับนำมาใช้ในการวิจัยเกี่ยวกับการวิเคราะห์ภาพใบหน้า โดยเฉพาะในบริบทของการตรวจจับใบหน้าปลอม



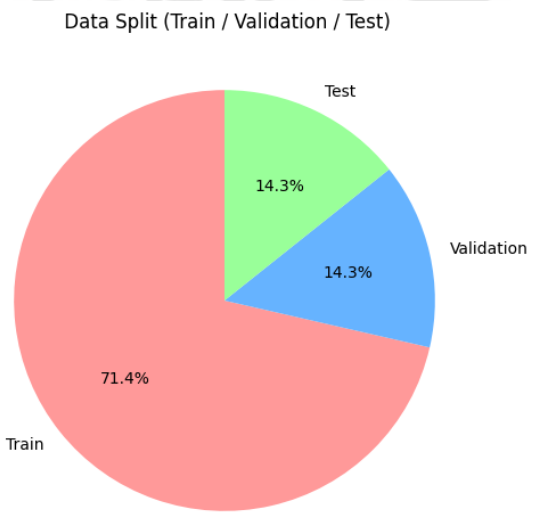
ภาพประกอบ 10 กราฟอัตราส่วนเชื้อชาติของภาพใบหน้าบุคคลจริงจากชุดข้อมูล 140K real VS fake faces ที่ใช้ทดสอบในงานวิจัย

ที่มา: https://github.com/NVlabs/ffhq-dataset

งานวิจัยเลือกใช้ข้อมูลเพียง 10% ของจำนวนภาพทั้งหมดเพื่อลดเวลาในการฝึกแบบจำลอง ใช้ข้อมูลภาพทั้งหมดเป็น 14,000 ภาพ แบ่งเป็นภาพใบหน้าจริงจำนวน 7,000 ภาพ และภาพใบหน้าปลอมจำนวน 7,000 ภาพ แบ่งข้อมูลออกเป็น ชุดข้อมูลฝึก (Training set) จำนวน 10,000 ภาพ , ชุดข้อมูลตรวจสอบ (Validation set) จำนวน 2,000 ภาพ , และชุดข้อมูลทดสอบ (Test set) จำนวน 2,000 ภาพ แสดงดังภาพประกอบ 11



ภาพประกอบ 11 จำนวนข้อมูลทั้งหมดที่ใช้ในงานวิจัย



ภาพประกอบ 12 กราฟแสดงอัตราส่วนของภาพที่ใช้ในการฝึกโมเดล

จากภาพประกอบ 12 แสดงอัตราส่วนของภาพที่ใช้ในการฝึกแบบจำลอง แบ่งข้อมูลออกเป็น

- ชุดข้อมูลฝึก (Training set) คิดเป็น 71.4% จากข้อมูลทั้งหมด
- ชุดข้อมูลตรวจสอบ (Validation set) 14.3% จากข้อมูลทั้งหมด
- ชุดข้อมูลทดสอบ (Test set) 14.3% จากข้อมูลทั้งหมด

จากภาพประกอบ 13 ตัวอย่าง ภาพจริงและภาพปลอม จากชุดข้อมูลฝึก (Training set)

ภาพประกอบ 13 : ตัวอย่างภาพจริงและภาพปลอมที่ได้จากชุดข้อมูลฝึก (Training set)

3.3 การเตรียมข้อมูลก่อนเข้าแบบจำลอง (Data preprocessing)

ขั้นตอนการเตรียมข้อมูลจากชุดข้อมูลที่ได้ โดยใช้โปรแกรม Google Colab ประเภท GPU และภาษาไพทอน (Python) พร้อมกับไลบรารี PyTorch ในงานวิจัยนี้ใช้การโหลดข้อมูลจากโฟลเดอร์ที่ได้จัดแบ่งเป็นชุดข้อมูล 3 โฟล์เดอร์ ได้แก่ ชุดข้อมูลฝึก (Training set), ชุดข้อมูลตรวจสอบ (Validation set) และ ชุดข้อมูลทดสอบ (Test set) ตามประเภทของภาพ คือ ภาพจริง และ ภาพปลอม กระบวนการเตรียมข้อมูลสำหรับแบบจำลองมีขั้นตอนดังนี้

3.3.1 การปรับขนาดภาพ (Image Resizing) ปรับขนาดของภาพทั้งหมดให้มีขนาด 224x224 พิกเซล เพื่อให้ภาพทั้งหมดมีขนาดที่เท่ากันก่อนป้อนเข้าสู่แบบจำลอง ซึ่งเป็นขนาดที่โมเดล เช่น ViT, DeiT และ Swin Transformer สามารถรับได้

3.3.2 การปรับภาพชุดข้อมูลให้หลากหลายขึ้น (Data Augmentation) ในข้อมูลชุดข้อมูลฝึก (training set) เพื่อช่วยเพิ่มความหลากหลายของข้อมูลและลดโอกาสที่แบบจำลองจะเกิด Overfitting การใช้เทคนิค Data Augmentation ช่วยให้แบบจำลองเรียนรู้ลักษณะของข้อมูลได้ดีขึ้น โดยในงานวิจัยนี้ ใช้การแปลงภาพดังนี้

- พลิกภาพแนวนอนแบบสุ่ม
- หมุนภาพแบบสุ่มไม่เกิน ±15 องศา
- ปรับค่าความสว่าง ความเข้ม สี และโทนสีของภาพแบบสุ่ม
- เลื่อนตำแหน่งของภาพในช่วง ±10%

3.3.3 การแปลงข้อมูลเป็น Tensor และการปรับค่าปกติ (Normalization) แปลงภาพเป็น Tensor เพื่อให้สามารถใช้กับโมเดลได้ ปรับค่าพิกเซลให้เป็นมาตรฐานโดยใช้ค่า Mean และ Standard Deviation ที่เหมาะสม

3.3.4 การกำหนดคลาสการจำแนก (Binary Classification) โมเดลถูกตั้งค่าให้ทำการจำแนกภาพออกเป็น 2 คลาส ได้แก่ ภาพจริง (Real, Label = 1) และภาพปลอม (Fake, Label = 0) โดยใช้การตั้งค่าแบบ binary classification ซึ่งโมเดลจะทำนายว่าแต่ละภาพที่ถูกป้อนเข้ามานั้นเป็นภาพจริงหรือภาพปลอม

3.4 การสร้างแบบจำลอง (Modeling)

ในงานวิจัยนี้ ได้มีการใช้โปรแกรม Google Colab ประเภท GPU ร่วมกับภาษาไพทอน (Python) เพื่อดำเนินการประมวลผลข้อมูลและสร้างแบบจำลองการจำแนกภาพ โดยใช้ไลบรารีที่สำคัญ ได้แก่ NumPy, Scikit-learn และ PyTorch ซึ่งเป็นเครื่องมือที่มีประสิทธิภาพในการจัดการกับข้อมูลขนาดใหญ่และการพัฒนาแบบจำลองการเรียนรู้เชิงลึก (Deep Learning) เทคนิคที่นำมาใช้ในงานวิจัยนี้คือ โครงข่ายประสาทเทียมที่ใช้เทคนิคทรานส์ฟอร์เมอร์ ซึ่งได้รับความนิยมอย่างแพร่หลายในการจำแนกภาพ เนื่องจากมีความสามารถในการจับลักษณะเชิงโครงสร้างของภาพได้ดี แบบจำลองที่เลือกใช้ในงานนี้ประกอบด้วย 3 แบบจำลอง ได้แก่ เช่น ViT , DeiT และ Swin Transformer ซึ่งทุกแบบจำลองได้ผ่านการฝึกฝนด้วยข้อมูลจาก ImageNet มาก่อน (Pre-trained Model) ทำให้สามารถนำมาใช้ในการจำแนกประเภทของภาพได้อย่างมีประสิทธิภาพ นอกจากนี้ งานวิจัยนี้ใช้เทคนิค Early Stopping เพื่อป้องกันปัญหา Overfitting ซึ่งเป็นปัญหาที่พบได้บ่อยในการฝึกแบบจำลองเชิงลึก โดยกำหนดค่า Patience = 5 หมายถึง หากค่าความสูญเสียของชุดข้อมูลตรวจสอบ (Validation Loss) ไม่ลดลงต่อเนื่องกันเป็น 5 รอบ การฝึกแบบจำลองจะถูกหยุดทันที เทคนิคนี้ช่วยให้สามารถเลือกโมเดลที่มีประสิทธิภาพสูงสุดโดยไม่ต้องฝึกมากเกินความจำเป็น ส่งผลให้ได้แบบจำลองที่มีความแม่นยำสูงและสามารถนำไปใช้งานได้อย่างมีประสิทธิภาพ ค่าพารามิเตอร์ที่ใช้ในการสร้างโมเดลจะถูกกำหนดไว้ในตารางที่แสดงด้านล่างดังนี้

ตาราง 1 แสดงพารามิเตอร์ที่ใช้ในการสร้างโมเดล

พารามิเตอร์	ค่า
Pre-trained Models	ImageNet-1k
Learning Rate	0.0001
Batch Size	32
Epoch	Early Stopping (patience = 5)
Optimizer	Adam
Loss Function	CrossEntropy

การสร้างแบบจำลองมีทั้งหมด 3 แบบจำลอง ดังนี้

3.4.1 Vision Transformer (ViT)

ในงานวิจัยส่วนแบบจำลอง ViT ที่เลือกใช้ คือ “vit_small_patch16_224” ในการทดสอบแบบจำลองซึ่งจะประกอบไปด้วยรายละเอียดของแต่ละชั้นดังนี้

- Patch Embedding แปลงภาพ 224x224 เป็น patches ขนาด 16x16
- Transformer Encoder ประกอบด้วย 12 layersของ Multi-Head Self-Attention และ Feedforward Network
- Output Layer การคำนวณผลลัพธ์สุดท้ายโดยการใช้ SoftMax Activation เพื่อจำแนกประเภทภาพ
-

3.4.2 Data-efficient Image Transformer (DeiT)

ในงานวิจัยส่วนแบบจำลองDeiT ที่เลือกใช้ คือ “deit_small_patch16_224” ในการทดสอบแบบจำลองซึ่งจะประกอบไปด้วยรายละเอียดของแต่ละชั้นดังนี้

- Patch Embedding แปลงภาพ 224x224 เป็น patches ขนาด 16x16
- Transformer Encoder ประกอบด้วย 12 layers ของ Multi-Head Self-Attention และ Feedforward Network
- Output Layer การคำนวณผลลัพธ์สุดท้ายโดยใช้ SoftMax Activation เพื่อจำแนกประเภทภาพ

3.4.3 Swin Transformer

ในงานวิจัยส่วนแบบจำลอง Swin Transformer เลือกใช้ คือ

“Swin_small_patch4_window7_224” ในการทดสอบซึ่งจะประกอบไปด้วยรายละเอียดของแต่ละชั้นดังนี้

- Patch Embedding แปลงภาพ 224x224 เป็น patches ขนาด 16x4
- Swin Transformer Blocks ประกอบด้วย 4 stages โดยแต่ละ stage ใช้ Window-based Multi-Head Attention ภายใน Windows ขนาด 7x7 และ Shifted Window Multi-Head Attention
- Output Layer การคำนวณผลลัพธ์สุดท้ายโดยใช้ SoftMax Activation เพื่อจำแนกประเภทภาพ

ตาราง 2 ค่าพารามิเตอร์ default setting ของทั้ง 3 แบบจำลอง

พารามิเตอร์	ViT	DeiT	Swin transformer
Image size	224	224	224
Patch size	16	16	4
Window size	-	-	7
num_channels	3	3	3
embed_dim / hidden_size	384	384	96
num_layers / depths	12	12	[2, 2, 18, 2]
num_attention_heads	6	6	[3, 6, 12, 24]
mlp_dim / intermediate_size	1536	1536	384
layer_norm_eps	1e-6	1e-6	1e-5

3.5 **การประเมินผล** (Evaluation)

การวิเคราะห์เปรียบเทียบที่ดำเนินการเห็นถึงประสิทธิภาพของโครงข่ายประสาทเทียมแบบ Transformer ในการแยกภาพ deepfake การประเมินผลการทำงานของโมเดลประกอบด้วยตัวชี้วัดที่แตกต่างกัน 4 ตัว ได้แก่ ความแม่นยำ (Accuracy), Precision, Recall และ F1-score

1) **ความแม่นยำ** (Accuracy) คำว่า "ความแม่นยำ" หมายถึงจำนวนกรณีที่ถูกจำแนกอย่างถูกต้อง คำนวณโดยใช้สูตรดังนี้

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (7)$$

TP (True Positives): จำนวนกรณีที่เป็นจริงและถูกจำแนกเป็นจริง

TN (True Negatives): จำนวนกรณีที่เป็นปลอมและถูกจำแนกเป็นปลอม

FP (False Positives): จำนวนกรณีที่เป็นปลอมแต่ถูกจำแนกเป็นจริง

FN (False Negatives): จำนวนกรณีที่เป็นจริงแต่ถูกจำแนกเป็นปลอม

2) Precision หมายถึง ความแม่นยำในการจำแนกประเภทบวก (positive) ของโมเดล โดยมีการวัดจากจำนวนกรณีที่ถูกจำแนกเป็นบวก (positive) ที่ถูกต้อง เมื่อเปรียบเทียบกับจำนวนกรณีทั้งหมดที่โมเดลจำแนกเป็นบวก (positive predictions) คำนวณโดยใช้สูตร

$$\text{Precision} = \frac{TP}{TP+FP} \quad (8)$$

3) Recall หมายถึง ความสามารถของโมเดลในการระบุกรณีที่เป็นบวก (positive) จริงในชุดข้อมูล โดยคำนวณจากจำนวนกรณีที่โมเดลจำแนกเป็นบวก (true positives) เมื่อเปรียบเทียบกับจำนวนกรณีที่เป็นบวกทั้งหมดในข้อมูล (true positives + false negatives) คำนวณโดยใช้สูตร

$$\text{Recall} = \frac{TP}{TP+FN} \quad (9)$$

4) F1-score คือ ดัชนีที่ใช้ในการประเมินประสิทธิภาพของโมเดลการจำแนกประเภท โดยเฉพาะในสถานการณ์ที่มีการไม่สมดุลระหว่างจำนวนตัวอย่างที่เป็นบวก (positive) และลบ

(negative) F1-score เป็นค่าเฉลี่ยแบบถ่วงน้ำหนักของ precision และ recall ซึ่งคำนึงถึงทั้ง false positives และ false negatives คำนวณโดยใช้สูตร

$$\mathrm{F1-score} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \quad (10)$$

ค่า F1-score จะมีค่าตั้งแต่ 0 ถึง 1

บทที่ 4
ผลการดำเนินงานวิจัย

ในการวิจัยเรื่องโครงข่ายประสาทเทียมทรานฟอร์เมอร์เพื่อการจำแนกภาพจริงและภาพปลอมผู้วิจัยดำเนินการวิจัยโดยศึกษาตามขั้นตอนและกระบวนการต่าง ๆ และได้มีการประเมินประสิทธิภาพของแบบจำลองให้เป็นไปตามวัตถุประสงค์ของการวิจัย ตามที่ตั้งเป้าหมายไว้ โดยมีแบบจำลองทั้งหมด 3 แบบจำลอง ได้แก่

1. สถาปัตยกรรมโครงข่ายประสาทเทียมทรานฟอร์เมอร์แบบจำลอง Vision Transformer (ViT) เพื่อการจำแนกภาพจริงและภาพปลอม
2. สถาปัตยกรรมโครงข่ายประสาทเทียมทรานฟอร์เมอร์แบบจำลอง Data-efficient Image Transformers (DeiT) เพื่อการจำแนกภาพจริงและภาพปลอม
3. สถาปัตยกรรมโครงข่ายประสาทเทียมทรานฟอร์เมอร์แบบจำลอง Swin Transformer เพื่อการจำแนกภาพจริงและภาพปลอม

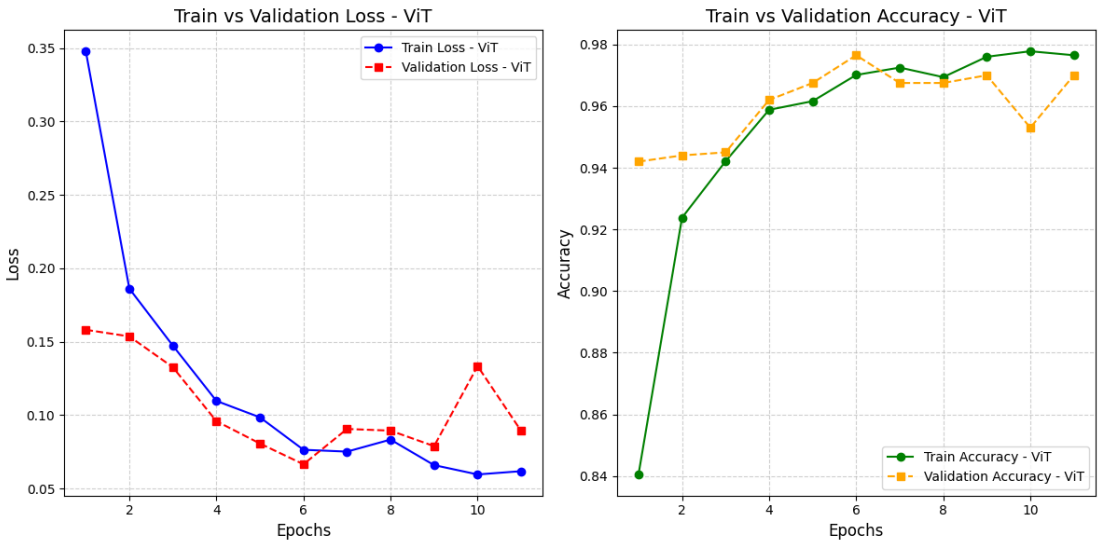
4.1 สถาปัตยกรรมโครงข่ายประสาทเทียมทรานฟอร์เมอร์แบบจำลอง Vision Transformer (ViT) เพื่อการจำแนกภาพจริงและภาพปลอม

จากภาพประกอบที่ 14 และตารางที่ 3 ซึ่งแสดงค่า Accuracy และ Loss ในแต่ละ Epoch พบว่าในการเรียนรู้บนชุดฝึกมีค่าความถูกต้องของชุดฝึก (Training Accuracy)ในช่วงต้นของการฝึกสอน Epoch 1 มีค่า 0.8405 และเพิ่มขึ้นอย่างต่อเนื่องใน Epoch ถัดมาจนมีค่าสูงสุดที่ใน Epoch 11 มีค่า 0.9765 ส่วนค่าความถูกต้องชุดข้อมูลตรวจสอบ (Validation Accuracy) สูงขึ้นเช่นกันโดยใน Epoch 1 มีค่า 0.9420 และสูงสุดที่ใน Epoch 6 มีค่า 0.9675 ก่อนที่จะมีแนวโน้มลดลงเล็กน้อยใน Epoch ถัดมา

ค่า Loss ในชุดข้อมูลฝึกสอน (Training Loss) ลดลงอย่างต่อเนื่องจาก Epoch 1 มีค่า 0.3478 จนเหลือค่าน้อยสุดใน Epoch 11 เท่ากับ 0.0618 แสดงถึงประสิทธิภาพที่ดีของแบบจำลอง ViT ในการเรียนรู้ชุดข้อมูลฝึกสอน ส่วนค่า Loss ของชุดข้อมูลตรวจสอบ (Validation Loss) มีแนวโน้มลดลงในช่วงแรก มีค่าต่ำสุดที่ Epoch 6 มีค่า 0.0665 แต่กลับเพิ่มขึ้นอย่างชัดเจนใน Epoch 10 เท่ากับ 0.1333 และ Epoch 11 มีค่า 0.0892

ในกระบวนการฝึกแบบจำลอง ViT มีการใช้ Early Stopping โดยกำหนดค่า Patience เท่ากับ 5 หากค่า Validation Loss ไม่ลดลงต่อเนื่องกัน 5 Epochs การฝึกจะหยุดลง จากภาพประกอบที่ 13 พบว่า Early Stopping Counter เริ่มทำงานที่ Epoch 7 และ จนถึง Epoch 11 ซึ่งเป็นจุดที่ถึงค่าที่กำหนดไว้ ส่งผลให้แบบจำลองหยุดการฝึกที่ Epoch 11 โดยอัตโนมัติ เพื่อป้องกันไม่ให้เกิด Overfitting มากขึ้น

จากการเปรียบเทียบค่า Accuracy และ Loss ในแต่ละรอบของการฝึกสอน พบว่าแบบจำลองมีการพัฒนาประสิทธิภาพอย่างต่อเนื่อง โดย Training Accuracy ในชุดข้อมูลฝึกสอนเพิ่มขึ้นอย่างรวดเร็วและมีค่าสูง ในขณะที่ Validation Accuracy ของชุดข้อมูลตรวจสอบแม้จะสูงแต่มีความผันผวนเล็กน้อย ส่วนด้านค่า Training Loss ของชุดข้อมูลฝึกสอนลดลงอย่างต่อเนื่องแสดงให้เห็นว่าแบบจำลองสามารถเรียนรู้ได้ดีขึ้น อย่างไรก็ตาม ค่า Validation Loss ที่เพิ่มน้อยขึ้นเล็กใน Epoch ท้ายๆ อาจบ่งชี้ถึงแนวโน้มของ Overfitting



ภาพประกอบ 14 กราฟแสดงค่า Loss และ ค่า Accuracy ระหว่างการเรียนรู้ของสถาปัตยกรรมโครงข่ายประสาทเทียมทรานฟอร์เมอร์แบบจำลอง ViT เพื่อการจำแนกภาพจริงและภาพปลอม

ตาราง 3 ค่า Loss และ ค่า Accuracy ของสถาปัตยกรรมโครงข่ายประสาทเทียมทรานฟอร์เมอร์แบบจำลอง ViT เพื่อการจำแนกภาพจริงและภาพปลอม

Epoch	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy
1	0.3478	0.8405	0.1581	0.9420
2	0.1861	0.9238	0.1536	0.9440
3	0.1473	0.9421	0.1326	0.9450
4	0.1097	0.9588	0.0958	0.9620
5	0.0985	0.9616	0.0806	0.9675
6	0.0764	0.9701	0.0665	0.9765
7	0.0751	0.9725	0.0905	0.9675
8	0.0833	0.9694	0.0894	0.9675
9	0.0659	0.9760	0.0788	0.9700
10	0.0596	0.9778	0.1333	0.9530
11	0.0618	0.9765	0.0892	0.9700

จากตาราง 4 ซึ่งแสดงคะแนนการวัดผลจากแบบจำลองด้วยชุดข้อมูลทดสอบ (Test set) ด้วยค่าความแม่นยำ (Accuracy), ค่าประมาณ Precision, Recall และ F1-Score ของสถาปัตยกรรมโครงข่ายประสาทเทียมทรานฟอร์เมอร์แบบจำลอง ViT เพื่อการจำแนกภาพจริงและภาพปลอม สามารถสรุปได้ว่า แบบจำลองนี้มีค่า Accuracy อยู่ที่ 0.9720 และสามารถจำแนกระหว่างภาพจริงและภาพปลอมได้ดี

ตาราง 4 คะแนนการวัดผลจากแบบจำลองด้วยชุดข้อมูลทดสอบของสถาปัตยกรรมโครงข่ายประสาทเทียมทรานฟอร์เมอร์แบบจำลอง ViT เพื่อการจำแนกภาพจริงและภาพปลอม

	precision	recall	f1-score	accuracy
Fake	0.96	0.98	0.97	
Real	0.98	0.96	0.97	0.97
weighted avg	0.97	0.97	0.97	

จากภาพประกอบ 15 แสดงผลลัพธ์ที่ได้จากการทำนายด้วยรูปแบบกราฟ Confusion Matrix จากชุดข้อมูลทดสอบ (Test set) ที่ใช้สถาปัตยกรรมโครงข่ายประสาทเทียมทรานฟอร์เมอร์แบบจำลอง ViT เพื่อการจำแนกภาพจริงและภาพปลอม ได้ค่าดังนี้

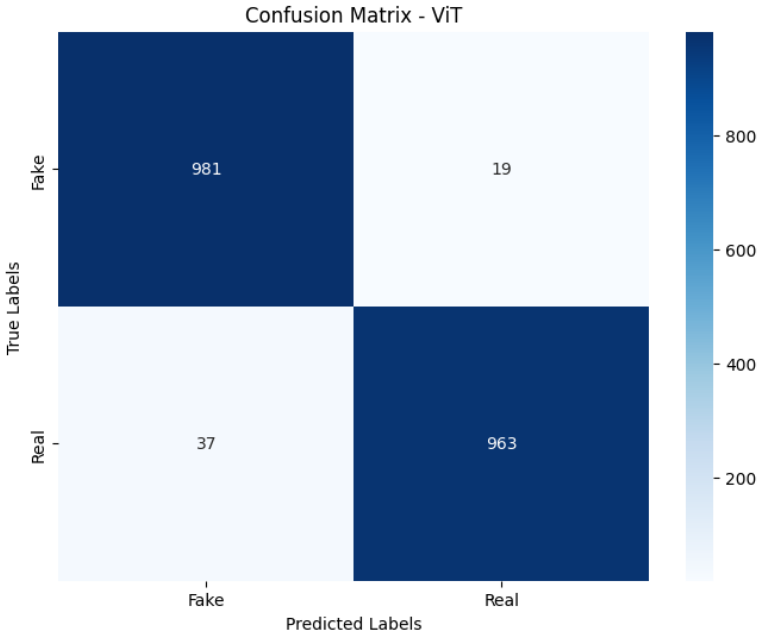
True Label: Fake ทำนายว่า Fake ซึ่งทำนายถูกต้องมีทั้งหมด 981 ภาพ

True Label: Fake ทำนายว่า Real ซึ่งทำนายผิดมีทั้งหมด 19 ภาพ

True Label: Real ทำนายว่า Real ซึ่งทำนายถูกต้องมีทั้งหมด 963 ภาพ

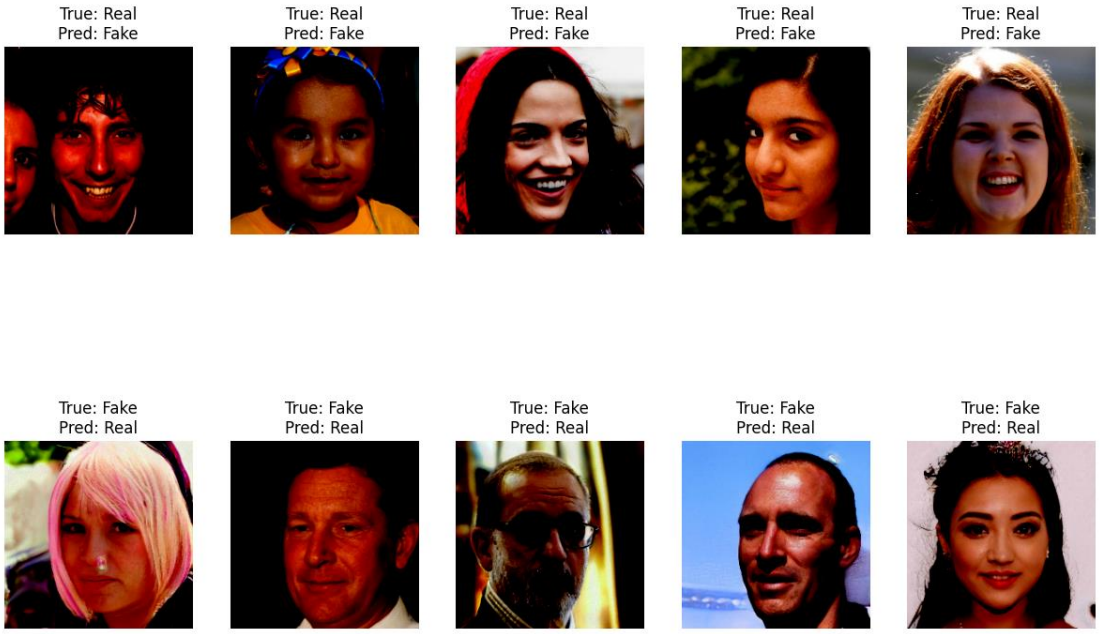
True Label: Real ทำนายว่า Fake ซึ่งทำนายผิดมีทั้งหมด 37 ภาพ

ผลลัพธ์แสดงให้เห็นว่าแบบจำลอง ViT มีความสามารถสูงในการตรวจจับภาพปลอม โดยมีค่า Recall ของคลาส Fake เท่ากับ 0.98 และ Real เท่ากับ 0.96 ซึ่งบ่งบอกถึงประสิทธิภาพของแบบจำลองในการระบุภาพปลอมและภาพจริงได้อย่างถูกต้อง เนื่องจากมีจำนวนภาพปลอมที่ถูกจำแนกผิดเป็นภาพจริงเพียง 19 ภาพ และ จำนวนภาพจริงที่ถูกจำแนกผิดเป็นภาพปลอมอยู่ที่ 37 ภาพ อัตราความผิดพลาด False Negative ต่ำกว่า False Positive ซึ่งหมายความว่าแบบจำลองมีแนวโน้มที่จะระบุภาพจริงว่าเป็นภาพปลอมมากกว่าการพลาดตรวจจับภาพปลอม



ภาพประกอบ 15 กราฟ Confusion Matrix แสดงแสดงผลลัพธ์ที่ได้จากการทำนายชุดทดสอบของสถาปัตยกรรมโครงข่ายประสาทเทียมทรานฟอร์เมอร์แบบจำลอง ViT เพื่อการจำแนกภาพจริงและภาพปลอม

จากภาพประกอบ 16 แสดงให้เห็นว่าภาพตัวอย่างที่แบบจำลอง ViT ทำนายผิดพลาดมีลักษณะผิดปกติ เช่น สีของภาพที่เพี้ยน ความมืดของภาพที่ทำให้มองไม่เห็นรายละเอียดสำคัญของใบหน้า เช่น ส่วนดวงตา ปาก และความไม่ชัดเจนของโครงสร้างใบหน้า ซึ่งอาจเกิดจากแสงเงาที่ไม่สม่ำเสมอ หรือการประมวลผลของภาพที่ผิดปกติ ส่งผลให้แบบจำลองไม่สามารถแยกแยะลักษณะของภาพจริงและปลอมได้อย่างแม่นยำ



ภาพประกอบ 16 ตัวอย่างภาพการทำนายผลผิด จากการชุดทดสอบของสถาปัตยกรรมโครงข่ายประสาทเทียมทรานฟอร์เมอร์แบบจำลอง ViT เพื่อการจำแนกภาพจริงและภาพปลอม

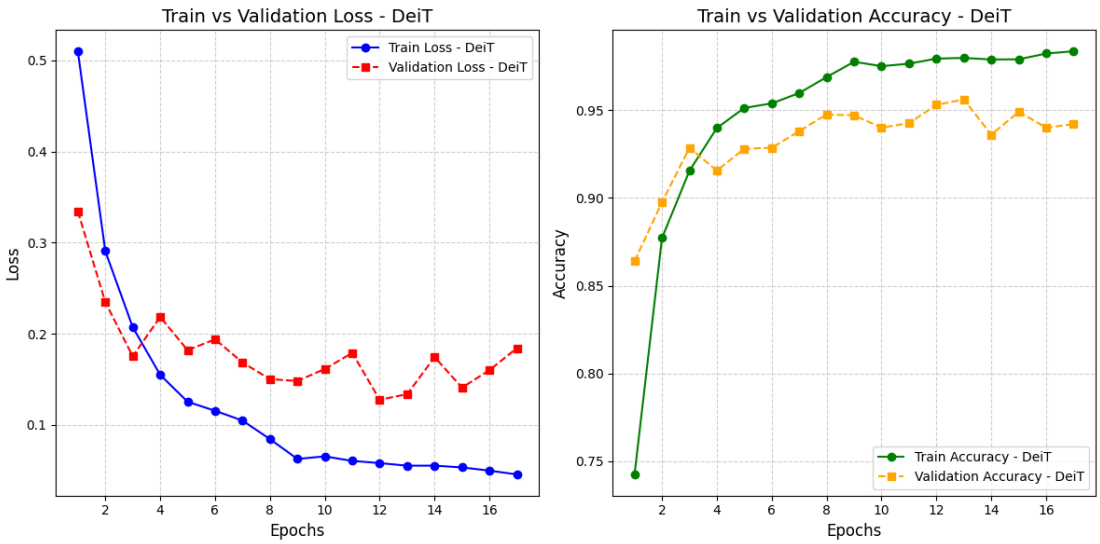
4.2 สถาปัตยกรรมโครงข่ายประสาทเทียมทรานฟอร์เมอร์แบบจำลอง Data-efficient Image Transformers (DeiT) เพื่อการจำแนกภาพจริงและภาพปลอม

จากภาพประกอบที่ 17 และตารางที่ 5 ซึ่งแสดงค่า Accuracy และ Loss ในแต่ละ Epoch พบว่าในการเรียนรู้บนชุดฝึกมีค่าความถูกต้องของชุดฝึก (Training Accuracy) ในช่วงเริ่มต้นของการฝึกสอน Epoch 1 มีค่า 0.7425 และเพิ่มขึ้นอย่างต่อเนื่องใน Epoch ถัดมาจนใน Epoch 17 มีค่า 0.9835 ส่วนค่าความถูกต้องชุดข้อมูลตรวจสอบ (Validation Accuracy) มีแนวโน้มเพิ่มขึ้นในช่วงต้นของการฝึก โดยใน Epoch 1 มีค่า 0.8640 และมีค่าสูงสุดใน Epoch 13 มีค่า 0.9560 ก่อนที่จะลดลงเล็กน้อยใน Epoch ถัดมา

ค่า Loss ในชุดข้อมูลฝึกสอน (Training Loss) ลดลงอย่างต่อเนื่องจาก ใน Epoch 1 มีค่า 0.5101 โดยใน Epoch 17 มีค่าเหลือเพียง 0.0457 แสดงถึงประสิทธิภาพที่ดีของแบบจำลอง DeiT ในการเรียนรู้ข้อมูลได้ดีขึ้นเรื่อย ๆ ส่วนค่า Loss ของชุดข้อมูลตรวจสอบ (Validation Loss) ใน Epoch 1 มีค่า 0.3344 มีแนวโน้มลดลงในช่วงแรกแต่มีแนวโน้มเพิ่มขึ้นในช่วงท้าย มีค่าต่ำสุดที่ Epoch 12 มีค่า 0.1275 และเพิ่มขึ้นเล็กน้อย แต่เพิ่มอย่างชัดเจนใน Epoch 14 จน Epoch 17 มีค่า 0.1843

ในกระบวนการฝึกแบบจำลอง DeiT มีการใช้ Early Stopping โดยกำหนดค่า Patience เท่ากับ 5 จากภาพประกอบที่ 16 พบว่า Early Stopping Counter เริ่มทำงานที่ Epoch 13 และจนถึง Epoch 17 ซึ่งเป็นจุดที่ถึงค่าที่กำหนดไว้ ส่งผลให้แบบจำลองหยุดการฝึกที่ Epoch 17 โดยอัตโนมัติ เพื่อป้องกันไม่ให้เกิด Overfitting มากขึ้น

จากการเปรียบเทียบค่า Accuracy และ Loss ในแต่ละรอบของการฝึกสอน พบว่าแบบจำลอง DeiT มีการพัฒนาประสิทธิภาพอย่างต่อเนื่อง โดย Training Accuracy ในชุดข้อมูลฝึกสอนเพิ่มขึ้นอย่างรวดเร็วและมีค่าสูงแสดงให้เห็นว่าแบบจำลอง DeiT สามารถเรียนรู้ข้อมูลได้ดีขึ้นเรื่อย ๆ ในขณะที่ Validation Accuracy ของชุดข้อมูลตรวจสอบมีค่าค่อนข้างสูง แต่เริ่มมีความผันผวนเล็กน้อยในช่วงท้ายของการฝึก ส่วนด้านค่า Training Loss ของชุดข้อมูลฝึกสอนลดลงอย่างต่อเนื่อง ซึ่งสามารถบ่งบอกว่าแบบจำลองสามารถเรียนรู้ลักษณะของข้อมูลฝึกสอนได้อย่างมีประสิทธิภาพ อย่างไรก็ตาม Validation Loss ที่เพิ่มขึ้นในบาง Epoch โดยเฉพาะหลังจาก Epoch ที่ 12 อาจเป็นสัญญาณของ Overfitting ซึ่งหมายความว่าโมเดลอาจเริ่มเรียนรายละเอียดของข้อมูลฝึกสอนมากเกินไป



ภาพประกอบ 17 กราฟแสดงค่า Loss และ ค่า Accuracy ระหว่างการเรียนรู้ของสถาปัตยกรรมโครงข่ายประสาทเทียมทรานฟอร์เมอร์แบบจำลอง DeiT เพื่อการจำแนกภาพจริงและภาพปลอม

ตาราง 5 ค่า Loss และ ค่า Accuracy ของสถาปัตยกรรมโครงข่ายประสาทเทียมทรานฟอร์เมอร์แบบจำลอง DeiT เพื่อการจำแนกภาพจริงและภาพปลอม

Epoch	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy
1	0.5101	0.7425	0.3344	0.8640
2	0.2907	0.8771	0.2349	0.8975
3	0.2074	0.9154	0.1753	0.9285
4	0.1550	0.9398	0.2185	0.9155
5	0.1254	0.9512	0.1820	0.9280
6	0.1156	0.9538	0.1940	0.9285
7	0.1048	0.9597	0.1685	0.9380
8	0.0845	0.9688	0.1503	0.9475
9	0.0627	0.9775	0.1482	0.9470
10	0.0655	0.9750	0.1614	0.9400
11	0.0605	0.9764	0.1791	0.9425
12	0.0581	0.9793	0.1276	0.9530

ตาราง 5 (ต่อ)

Epoch	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy
13	0.0553	0.9797	0.1338	0.9560
14	0.0553	0.9788	0.1744	0.9360
15	0.0534	0.9789	0.1410	0.9490
16	0.0498	0.9822	0.1601	0.9400
17	0.0457	0.9835	0.1843	0.9420

จากตาราง 6 ซึ่งแสดงคะแนนการวัดผลจากแบบจำลองด้วยชุดข้อมูลทดสอบ (Test set) ด้วยค่าความแม่นยำ (Accuracy), ค่าประมาณ Precision, Recall และ F1-Score ของสถาปัตยกรรมโครงข่ายประสาทเทียมทรานฟอร์เมอร์แบบจำลอง DeiT เพื่อการจำแนกภาพจริงและภาพปลอม สามารถสรุปได้ว่า แบบจำลองนี้มีค่า Accuracy อยู่ที่ 0.9525 และสามารถจำแนกระหว่างภาพจริงและภาพปลอมได้ดี

ตาราง 6 คะแนนการวัดผลจากแบบจำลองด้วยชุดข้อมูลทดสอบ ของสถาปัตยกรรมโครงข่ายประสาทเทียมทรานฟอร์เมอร์แบบจำลอง DeiT เพื่อการจำแนกภาพจริงและภาพปลอม

	precision	recall	f1-score	accuracy
Fake	0.94	0.97	0.95	0.95
Real	0.97	0.94	0.95	
weighted avg	0.95	0.95	0.95	

จากภาพประกอบ 18 แสดงผลลัพธ์ที่ได้จากการทำนายด้วยรูปแบบกราฟ Confusion Matrixจากชุดข้อมูลทดสอบ (Test set) ที่ใช้สถาปัตยกรรมโครงข่ายประสาทเทียมทรานฟอร์เมอร์แบบจำลอง DeiT เพื่อการจำแนกภาพจริงและภาพปลอม ได้ค่าดังนี้

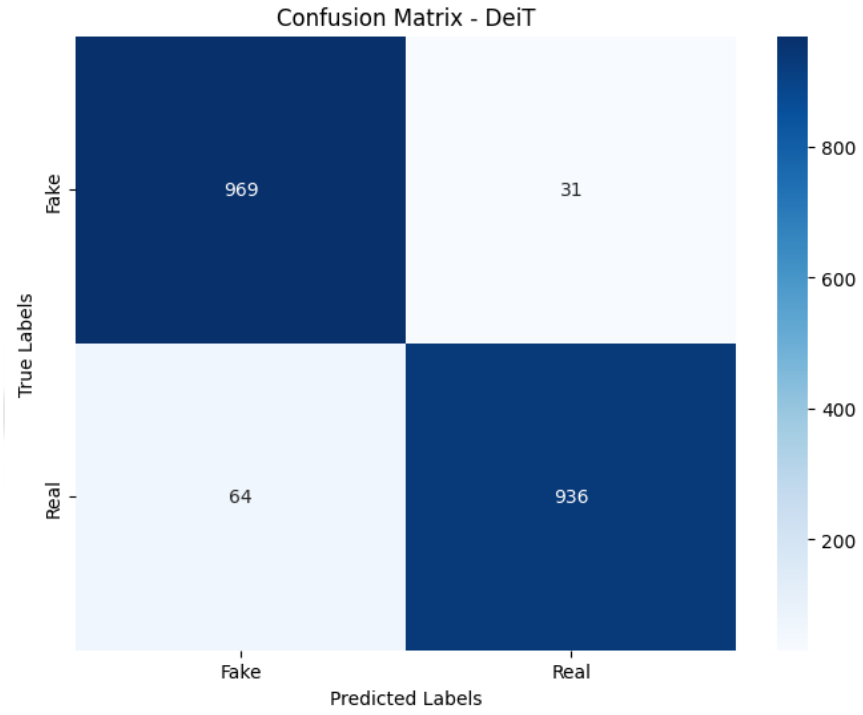
True Label: Fake ทำนายว่า Fake ซึ่งทำนายถูกต้องมีทั้งหมด 969 ภาพ

True Label: Fake ทำนายว่า Real ซึ่งทำนายผิดมีทั้งหมด 31 ภาพ

True Label: Real ทำนายว่า Real ซึ่งทำนายถูกต้องมีทั้งหมด 936 ภาพ

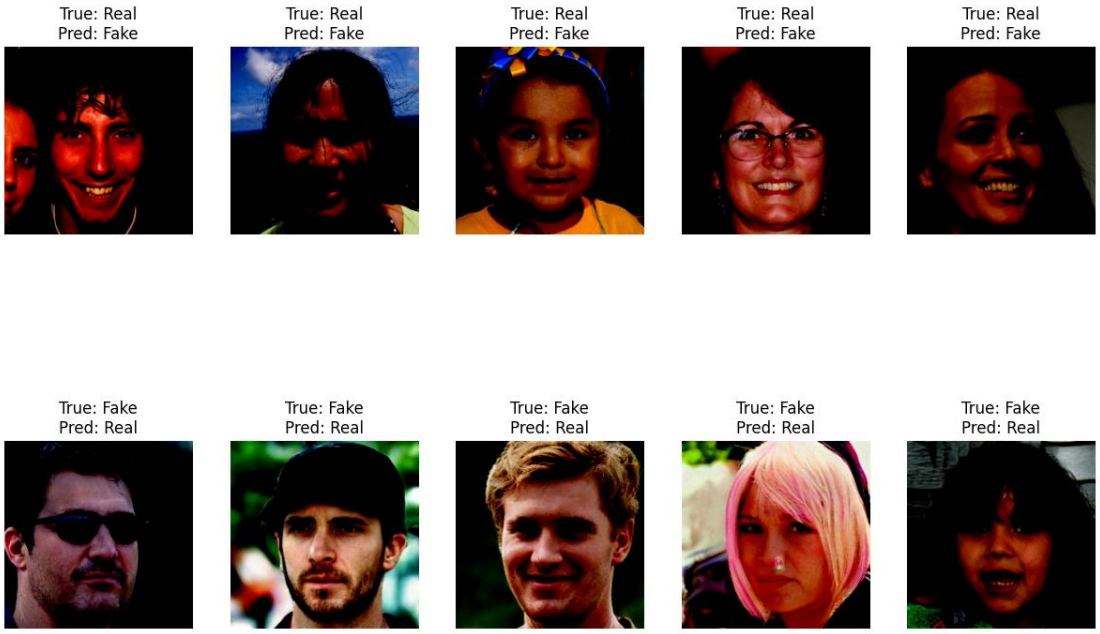
True Label: Real ทำนายว่า Fake ซึ่งทำนายผิดมีทั้งหมด 64 ภาพ

ผลลัพธ์แสดงให้เห็นว่า แบบจำลอง DeiT มีความสามารถสูงในการตรวจจับภาพปลอม โดยมีค่า Recall ของคลาสภาพปลอมที่ทำนายถูก เท่ากับ 0.97 ซึ่งสะท้อนถึงประสิทธิภาพของแบบจำลองในการระบุภาพปลอมได้อย่างถูกต้อง เนื่องจากจำนวนภาพปลอมที่ถูกจำแนกผิดเป็นภาพจริงมีเพียง 31 ภาพ แต่อย่างไรก็ตาม จำนวนภาพจริงที่ถูกจำแนกผิดเป็นภาพปลอมอยู่ที่ 64 ภาพ ซึ่ง False Positive สูงกว่า False Negative หมายความว่าแบบจำลองมีแนวโน้มที่จะจำแนกภาพจริงผิดเป็นภาพปลอมมากกว่าจำแนกภาพปลอมผิดเป็นภาพจริง



ภาพประกอบ 18 กราฟ Confusion Matrix แสดงแสดงผลลัพธ์ที่ได้จากการทำนายชุดทดสอบของสถาปัตยกรรมโครงข่ายประสาทเทียมทรานฟอร์เมอร์แบบจำลอง DeiT เพื่อการจำแนกภาพจริงและภาพปลอม

จากภาพประกอบ 19 แสดงให้เห็นว่าภาพตัวอย่างที่แบบจำลอง DeiT ทำนายผิดพลาดมีลักษณะผิดปกติ เช่น สีของภาพที่เพี้ยน ความมืดของภาพที่ทำให้มองไม่เห็นรายละเอียดสำคัญของใบหน้า เช่น ส่วนดวงตา ปาก และความไม่ชัดเจนของโครงสร้างใบหน้า ซึ่งอาจเกิดจากแสงเงาที่ไม่สม่ำเสมอ หรือการประมวลผลของภาพที่ผิดปกติ ส่งผลให้แบบจำลองไม่สามารถแยกแยะลักษณะของภาพจริงและปลอมได้อย่างแม่นยำ



ภาพประกอบ 19 ตัวอย่างภาพการทำนายผลผิด จากการชุดทดสอบของสถาปัตยกรรมโครงข่ายประสาทเทียมทรานฟอร์เมอร์แบบจำลอง DeiT เพื่อการจำแนกภาพจริงและภาพปลอม

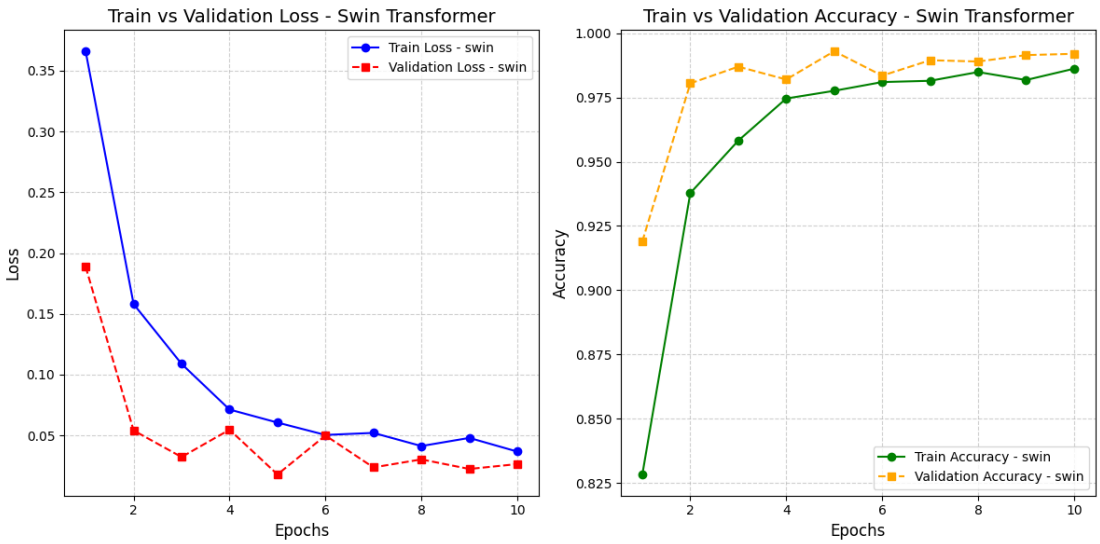
4.3 สถาปัตยกรรมโครงข่ายประสาทเทียมทรานฟอร์เมอร์แบบจำลอง Swin Transformer เพื่อการจำแนกภาพจริงและภาพปลอม

จากภาพประกอบที่ 20 และตารางที่ 7 ซึ่งแสดงค่า Accuracy และ Loss ในแต่ละ Epoch พบว่าในการเรียนรู้บนชุดฝึกมีค่าความถูกต้องของชุดฝึก (Training Accuracy) ในช่วงเริ่มต้นของการฝึกสอน Epoch 1 มีค่า 0.8283 และเพิ่มขึ้นอย่างต่อเนื่องใน Epoch ถัดมาจนใน Epoch 10 มีค่า 0.9862 ส่วนค่าความถูกต้องชุดข้อมูลตรวจสอบ (Validation Accuracy) มีแนวโน้มเพิ่มขึ้นเรื่อย ๆในช่วงต้นของการฝึก โดยใน Epoch 1 มีค่า 0.9190 โดยมีค่าสูงสุดใน Epoch 5 มีค่า 0.9930 และมีค่าลดลงเล็กน้อยส่วนใน Epoch 10 มีค่า 0.9920

ค่า Loss ในชุดข้อมูลฝึกสอน (Training Loss) ลดลงอย่างต่อเนื่องจาก ใน Epoch 1 มีค่า 0.3660 โดยใน Epoch 10 มีค่าเหลือเพียง 0.0368 แสดงถึงประสิทธิภาพที่ดีของแบบจำลอง Swin Transformer ในการเรียนรู้ข้อมูลได้ดีขึ้นเรื่อย ๆ ส่วนค่า Loss ของชุดข้อมูลตรวจสอบ (Validation Loss) ลดลงจาก Epoch 1 มีค่า 0.1888 ลดจนเรื่อยๆมีค่าต่ำที่สุดใน Epoch 5 มีค่า 0.0179 แสดงว่าแบบจำลอง Swin Transformer มีความเสถียรสูง แม้จะมีความผันผวนของ Validation Loss เล็กน้อยในช่วงเริ่มต้นเทรน

ในกระบวนการฝึกแบบจำลอง Swin Transformer มีการใช้ Early Stopping โดยกำหนดค่า Patience เท่ากับ 5 จากภาพประกอบที่ 19 พบว่า Early Stopping Counter เริ่มทำงานที่ Epoch 6 และ จนถึง Epoch 10 ซึ่งเป็นจุดที่ถึงค่าที่กำหนดไว้ ส่งผลให้แบบจำลองหยุดการฝึกที่ Epoch 10 โดยอัตโนมัติ เพื่อป้องกันไม่ให้เกิด Overfitting มากขึ้น

จากการเปรียบเทียบค่า Accuracy และ Loss ในแต่ละรอบของการฝึกสอน พบว่าแบบจำลอง Swin Transformer มีการพัฒนาประสิทธิภาพอย่างต่อเนื่อง โดย Training Accuracy ของชุดข้อมูลฝึกสอนเพิ่มขึ้นอย่างรวดเร็วและมีค่าคงที่ในระดับสูง แสดงให้เห็นว่าแบบจำลองสามารถเรียนรู้ข้อมูลได้อย่างมีประสิทธิภาพ ในขณะที่ Validation Accuracy ของชุดข้อมูลตรวจสอบมีค่าค่อนข้างสูง โดยแตะระดับ 99% ตั้งแต่ Epoch 5 อย่างไรก็ตาม มีความผันผวนเล็กน้อยช่วงของการฝึก ด้านค่า Training Loss ของชุดข้อมูลฝึกสอนลดลงอย่างต่อเนื่อง ซึ่งบ่งชี้ว่าโมเดลสามารถเรียนรู้โครงสร้างของข้อมูลได้ดีขึ้นเรื่อย ๆ อย่างไรก็ตาม Validation Loss มีแนวโน้มลดลง ซึ่งอาจจะผันผวนเล็กน้อยในบาง Epoch โดยเฉพาะ Epoch ที่ 4 และ 6



ภาพประกอบ 20 กราฟแสดงค่า Loss และ ค่า Accuracy ระหว่างการเรียนรู้ของสถาปัตยกรรมโครงข่ายประสาทเทียมทรานฟอร์เมอร์แบบจำลอง Swin Transformer เพื่อการจำแนกภาพจริงและภาพปลอม

ตาราง 7 ค่า Loss และ ค่า Accuracy ของสถาปัตยกรรมโครงข่ายประสาทเทียมทรานฟอร์เมอร์แบบจำลอง Swin Transformer เพื่อการจำแนกภาพจริงและภาพปลอม

Epoch	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy
1	0.3660	0.8283	0.1888	0.9190
2	0.1580	0.9378	0.0541	0.9805
3	0.1089	0.9582	0.0323	0.9870
4	0.0713	0.9746	0.0547	0.9820
5	0.0606	0.9776	0.0179	0.9930
6	0.0504	0.9810	0.0500	0.9835
7	0.0521	0.9815	0.0237	0.9895
8	0.0413	0.9849	0.0303	0.9890
9	0.0480	0.9818	0.0224	0.9915
10	0.0368	0.9862	0.0265	0.9920

จากตาราง 8 ซึ่งแสดงคะแนนการวัดผลจากแบบจำลองด้วยชุดข้อมูลทดสอบ (Test set) ด้วยค่าความแม่นยำ (Accuracy), ค่าประมาณ Precision, Recall และ F1-Score ของสถาปัตยกรรมโครงข่ายประสาทเทียมทรานฟอร์เมอร์แบบจำลอง Swin Transformer เพื่อการจำแนกภาพจริงและภาพปลอม สามารถสรุปได้ว่า แบบจำลองนี้มีค่า Accuracy อยู่ที่ 0.9915 สามารถจำแนกระหว่างภาพจริงและภาพปลอมได้อย่างมีประสิทธิภาพสูง

ตาราง 8 คะแนนการวัดผลจากแบบจำลองด้วยชุดข้อมูลทดสอบ ของสถาปัตยกรรมโครงข่ายประสาทเทียมทรานฟอร์เมอร์แบบจำลอง Swin Transformer เพื่อการจำแนกภาพจริงและภาพปลอม

	precision	recall	f1-score	accuracy
Fake	0.99	0.99	0.99	0.99
Real	0.99	0.99	0.99	
weighted avg	0.99	0.99	0.99	

จากภาพประกอบ 21 แสดงผลลัพธ์ที่ได้จากการทำนายด้วยรูปแบบกราฟ Confusion Matrixจากชุดข้อมูลทดสอบ (Test set) ที่ใช้สถาปัตยกรรมโครงข่ายประสาทเทียมทรานฟอร์เมอร์แบบจำลอง Swin Transformer เพื่อการจำแนกภาพจริงและภาพปลอม ได้ค่าดังนี้

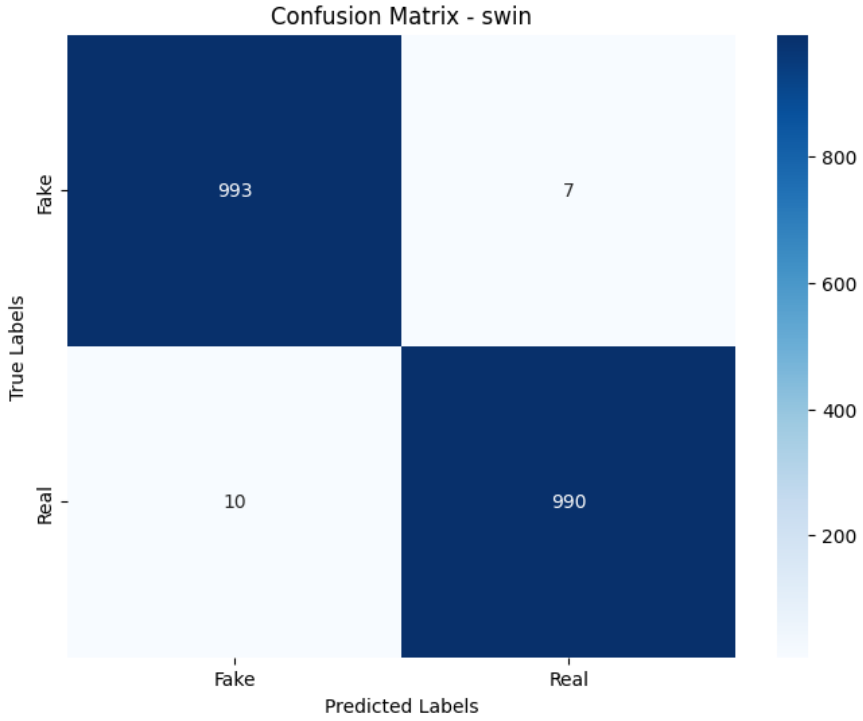
True Label: Fake ทำนายว่า Fake ซึ่งทำนายถูกต้องมีทั้งหมด 993 ภาพ

True Label: Fake ทำนายว่า Real ซึ่งทำนายผิดมีทั้งหมด 7 ภาพ

True Label: Real ทำนายว่า Real ซึ่งทำนายถูกต้องมีทั้งหมด 990 ภาพ

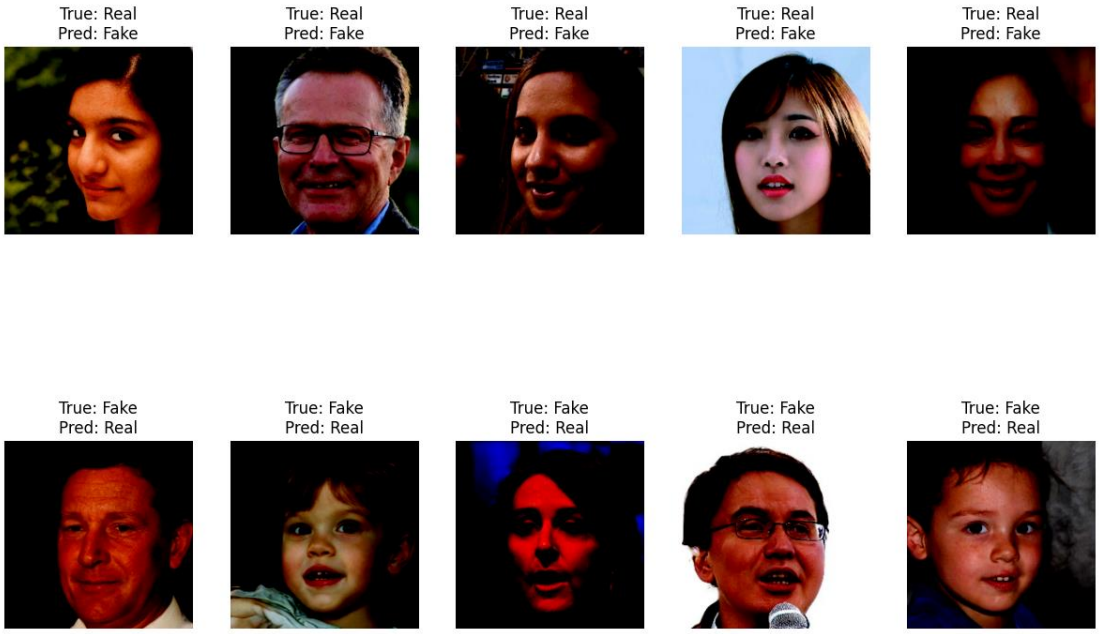
True Label: Real ทำนายว่า Fake ซึ่งทำนายผิดมีทั้งหมด 10 ภาพ

ผลลัพธ์แสดงให้เห็นว่า แบบจำลอง Swin Transformer สามารถตรวจจับภาพปลอมได้อย่างแม่นยำ มีค่า Recall ของคลาส Fake สูงมาก เนื่องจากมีภาพปลอมที่ทำนายผิดเพียง 7 ภาพ ขณะเดียวกัน จำนวนภาพจริงที่ถูกทำนายผิดเป็นภาพปลอมอยู่ที่ 10 ภาพ ซึ่งบ่งชี้ว่าแบบจำลองมีแนวโน้มที่จะทำนายผิดในทิศทางนี้เล็กน้อย โดยรวมแล้ว Swin Transformer แสดงศักยภาพสูงในการจำแนกภาพจริงและภาพปลอม โดยสามารถลดข้อผิดพลาดให้น้อยลง ทำให้เป็นแบบจำลองที่มีความแม่นยำและเสถียรภาพสูงในการใช้งาน



ภาพประกอบ 21 กราฟ Confusion Matrix แสดงแสดงผลลัพธ์ที่ได้จากการทำนายชุดทดสอบของสถาปัตยกรรมโครงข่ายประสาทเทียมทรานฟอร์เมอร์แบบจำลอง Swin Transformer เพื่อการจำแนกภาพจริงและภาพปลอม

จากภาพประกอบ 22 แสดงให้เห็นว่าภาพตัวอย่างที่แบบจำลอง Swin Transformer ทำนายผิดพลาดมีลักษณะผิดปกติ เช่น สีของภาพที่เพี้ยน ความมืดของภาพที่ทำให้มองไม่เห็นรายละเอียดสำคัญของใบหน้า เช่น ส่วนดวงตา ปาก และความไม่ชัดเจนของโครงสร้างใบหน้า ซึ่งอาจเกิดจากแสงเงาที่ไม่สม่ำเสมอ หรือการประมวลผลของภาพที่ผิดปกติ ส่งผลให้แบบจำลองไม่สามารถแยกแยะลักษณะของภาพจริงและปลอมได้อย่างแม่นยำ



ภาพประกอบ 22 ตัวอย่างภาพการทำนายผลผิด จากการชุดทดสอบของสถาปัตยกรรมโครงข่ายประสาทเทียมทรานฟอร์เมอร์แบบจำลอง Swin Transformer เพื่อการจำแนกภาพจริงและภาพปลอม

บทที่ 5
สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

ในการวิจัยเรื่องได้วัดประสิทธิภาพแบบโครงข่ายประสาทเทียมทรานฟอร์เมอร์เพื่อการจำแนกภาพจริงและภาพปลอมแบบจำลองทั้ง 3 ประเภท ได้แก่ แบบจำลอง Vision Transformer (ViT) , Data-efficient Image Transformer (DeiT) และ Swin Transformer เพื่อนำมาเปรียบเทียบและสรุป โดยสามารถแบ่งหัวข้อสรุปผลได้ดังนี้

1. สรุปผลการวิจัย
2. อภิปรายผลการวิจัย
3. ข้อเสนอแนะ

5.1 สรุปผลการวิจัย

จากการศึกษาและเปรียบเทียบประสิทธิภาพของแบบจำลองโครงข่ายประสาทเทียมทรานฟอร์เมอร์ในการจำแนกภาพจริงและภาพปลอมในงานวิจัยนี้ทั้งแบบจำลอง 3 ประเภท ได้แก่ Vision Transformer (ViT), Data-efficient Image Transformer (DeiT) และ Swin Transformer เพื่อประเมินความสามารถแบบจำลองในการจำแนกภาพจริงและภาพปลอมได้อย่างมีประสิทธิภาพสูงสุดโดยการเปรียบเทียบประสิทธิภาพของแต่ละแบบจำลอง ดังตาราง 9 และสามารถอธิบายสรุปผลการดำเนินการวิจัยได้ดังนี้

ประสิทธิภาพของแบบจำลองเปรียบเทียบค่าความแม่นยำ (Accuracy) ของโครงข่ายประสาทเทียมทรานฟอร์เมอร์จำลองทั้ง 3 แบบจำลองในชุดข้อทดสอบที่จำแนกข้อมูลภาพจริงและภาพปลอมจากผลการทดลอง พบว่าแบบจำลอง Swin Transformer เป็นแบบจำลองที่มีค่าความแม่นยำสูงสุดที่ 0.9915 แสดงให้เห็นถึงความสามารถในการเรียนรู้คุณลักษณะของภาพได้อย่างมีประสิทธิภาพมากที่สุด รองลงมาคือแบบจำลอง ViT มีค่าความแม่นยำ 0.9720 และ แบบจำลอง DeiT มีประสิทธิภาพของแบบจำลองต่ำที่สุด มีค่าความแม่นยำที่ 0.9525 เมื่อทำการเปรียบเทียบประสิทธิภาพของแบบจำลองทั้ง 3 แบบจำลอง

ตาราง 9 การเปรียบเทียบค่าความแม่นยำ (Accuracy) และ Weight Average ของโครงข่ายประสาทเทียมทรานฟอร์เมอร์จำลองทั้ง 3 แบบจำลองในชุดข้อทดสอบที่จำแนกข้อมูลภาพจริงและภาพปลอม

แบบจำลอง	Weight Average Precision	Weight Average Recall	Weight Average F1-Score	Accuracy
ViT	0.97	0.97	0.97	0.9720
DeiT	0.95	0.95	0.95	0.9525
Swin	0.99	0.99	0.99	0.9915

5.2 อภิปรายผลการวิจัย

เปรียบเทียบผลลัพธ์ของแบบจำลองที่นำเสนอในงานวิจัยนี้กับผลงานของ Sharma J และคณะ(30) ซึ่งใช้ชุดข้อมูลทดสอบเดียวกัน ได้แก่ 140k Real and Fake Faces พบว่างานวิจัยของ Sharma J และคณะ ที่เลือกใช้แบบจำลองเทคนิคโครงข่ายประสาทเทียมแบบคอนโวลูชัน (CNN) ได้แก่ ResNet50 และ VGG16 ให้ค่าความแม่นยำได้เท่ากับ 0.9398 และ 0.8663 ตามลำดับ ในขณะที่แบบจำลองที่นำเสนอในงานวิจัยนี้ ซึ่งประยุกต์ใช้กับเทคนิคโครงข่ายประสาทเทียมแบบ Transformer ได้แก่ ViT, DeiT และ Swin Transformer สามารถให้ผลลัพธ์ที่เหนือกว่าอย่างมีนัยสำคัญ โดยให้ค่าความแม่นยำเท่ากับ 0.9915, 0.9720 และ 0.9525 ตามลำดับ สะท้อนให้เห็นถึงศักยภาพของแบบจำลองในกลุ่ม Transformer ที่มีการเรียนรู้คุณลักษณะเชิงลึกของภาพผ่านกลไก Self-Attention ซึ่งสามารถตรวจจับรายละเอียดที่ละเอียดอ่อนและความผิดปกติที่แฝงอยู่ในภาพปลอมได้อย่างแม่นยำยิ่งขึ้น ดังนั้นการเลือกใช้แบบจำลองในกลุ่ม Transformer จึงเป็นตัวเลือกที่เหมาะสมสำหรับงานวิจัยที่ต้องการแยกภาพจริงกับภาพปลอม

จากการศึกษาประสิทธิภาพของโครงข่ายประสาทเทียมทรานฟอร์เมอร์ในการจำแนกภาพจริงและภาพปลอม โดยเปรียบเทียบแบบจำลอง ViT, DeiT และ Swin Transformer พบว่าแบบจำลองแต่ละตัวมีประสิทธิภาพที่แตกต่างกันตามโครงสร้างและหลักการทำงานของแบบจำลอง

การเปรียบเทียบประสิทธิภาพของแบบจำลอง จากผลการทดลองแสดงให้เห็นว่า Swin Transformer เป็นแบบจำลองที่มีค่าความแม่นยำสูงสุดที่ 0.9915 และมีค่า Precision, Recall และ F1-Score เท่ากับ 0.99 สะท้อนถึงความสามารถในการเรียนรู้คุณลักษณะของภาพได้อย่างมีประสิทธิภาพที่สุดรองลงมาคือแบบจำลอง ViT มีค่าความแม่นยำ 0.9720 โดยมีค่า Precision, Recall และ F1-Score อยู่ที่ 0.97 ซึ่งแม้ว่าจะมีประสิทธิภาพดี แต่ยังมีข้อจำกัดในการเรียนรู้ความสัมพันธ์เชิงพื้นที่ของภาพ และ DeiT มีค่าความแม่นยำต่ำที่สุดที่ 0.9525 และมีค่า Precision, Recall และ F1-Score เท่ากับ 0.95 เนื่องจากยังคงใช้โครงสร้างแบบเดียวกับ ViT โดยไม่มีการเพิ่มกลไกที่ช่วยให้สามารถจับรายละเอียดของภาพได้ดีขึ้น จากผลการทดลองพบว่าแบบจำลอง ViT และ DeiT มีประสิทธิภาพไม่ดีเท่ากับ Swin Transformer แม้แบบจำลองทั้ง 3 แบบจำลองจะมีการเทรนโดยใช้ Early Stopping เพื่อป้องกันปัญหา Overfitting แต่ยังเกิดยังการ Overfitting ขึ้นสาเหตุมาจากแบบจำลอง ViT และ DeiT ต้องข้อมูลขนาดใหญ่ในการเทรนเพื่อป้องกันการจดจำข้อมูลมากเกินไป เนื่องจากแบบจำลองมีพารามิเตอร์จำนวนมาก และหากชุดข้อมูลฝึก (Training Set) มีขนาดไม่ใหญ่พอ แบบจำลองจะมีแนวโน้ม จดจำแทนที่จะเรียนรู้ทำให้สามารถทำนายได้ดีเฉพาะบนชุดข้อมูลฝึก แต่ไม่สามารถ Generalize ไปยังข้อมูลใหม่ได้ดีพออีกทั้งความซับซ้อนของแบบจำลองใช้ Self-Attention แบบ Global Attention ซึ่งคำนึงถึงทุกส่วนของภาพพร้อมกันอาจไม่เหมาะสมกับการเรียนรู้โครงสร้างภาพที่ซับซ้อน ขณะที่ Swin Transformer แสดงศักยภาพได้โดดเด่นที่สุดในบรรดาแบบจำลองที่ใช้ในงานวิจัยนี้ เนื่องจากอาศัยกลไก Window-based Self-Attention ร่วมกับโครงสร้างการเรียนรู้แบบลำดับชั้น (Hierarchical Learning) ซึ่งมีความคล้ายคลึงกับโครงข่ายประสาทเทียมแบบคอนโวลูชัน กล่าวคือเริ่มต้นจากการเรียนรู้คุณลักษณะในระดับต่ำ (low-level features) และค่อย ๆ สังเคราะห์เป็นคุณลักษณะที่ซับซ้อนมากยิ่งขึ้นในแต่ละระดับชั้นของแบบจำลอง ส่งผลให้สามารถเข้าใจได้ทั้งคุณลักษณะเฉพาะจุดและบริบทโดยรวมของภาพได้อย่างมีประสิทธิภาพ จึงเหมาะสมอย่างยิ่งสำหรับการประยุกต์ใช้ในงานตรวจจับภาพปลอมที่มีความซับซ้อนในเชิงโครงสร้างของภาพ

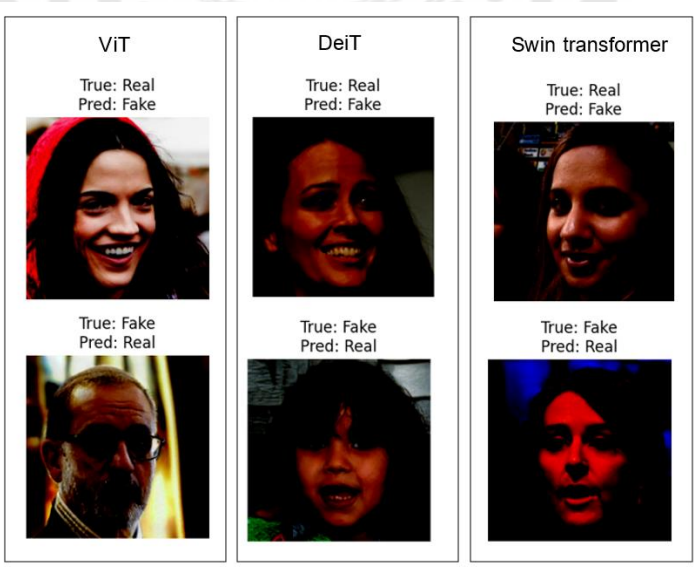
จากการวิเคราะห์ข้างต้น สามารถสรุปได้ว่า Swin Transformer เป็นแบบจำลองที่เหมาะสมที่สุดสำหรับการจำแนกภาพปลอมในงานวิจัยนี้ เนื่องจากสามารถเรียนรู้คุณลักษณะเชิงลึกของภาพได้อย่างมีประสิทธิภาพโดยไม่เกิดปัญหา Overfitting ขณะที่แบบจำลอง ViT และ DeiT แม้จะให้ผลลัพธ์ในระดับที่ดี แต่ยังพบข้อจำกัดในการประยุกต์ใช้กับข้อมูลที่มีความหลากหลายหรือมีโครงสร้างซับซ้อน โดยเฉพาะในกรณีที่มีข้อมูลฝึกจำนวนน้อย ซึ่งอาจส่งผลให้เกิด Overfitting หรือประสิทธิภาพลดลง ดังนั้น เพื่อเพิ่มความสามารถในการจำแนกภาพ ควรมี

การปรับปรุงเพิ่มเติม เช่น การใช้เทคนิค การใช้ Data Augmentation ที่มีความซับซ้อนมากขึ้น หรือการลดจำนวนพารามิเตอร์ เพื่อเพิ่มความสามารถในการประยุกต์ใช้กับข้อมูลที่หลากหลายยิ่งขึ้น

จากภาพประกอบ 23 แสดงผลลัพธ์ heatmap ของ SHAP value ภาพตัวอย่างจากการทำนายด้วยแบบจำลอง Swin transformer ซึ่งใช้เพื่ออธิบายว่าแบบจำลองพิจารณาส่วนใดของภาพในการตัดสินว่าเป็นภาพปลอมหรือภาพจริง ส่วนแรก ภาพด้านบน มีส่วนที่ค่า SHAP value เป็นบวก (แสดงด้วยสีแดง) โดยภาพมีสีแดงเข้มที่ดวงตา ขอบหน้า และหน้าผาก ค่า SHAP value เป็นลบ (แสดงด้วยสีน้ำเงิน) มีบางจุดสีน้ำเงินบริเวณข้างแก้ม ซึ่งเป็นบริเวณที่แบบจำลองคิดว่า ไม่เหมือนภาพปลอม จุดสีแดงแหล่งนี้นับเป็น feature สำคัญที่ทำให้แบบจำลองทำนายว่าภาพนี้เป็นภาพปลอม ความเข้มของสียิ่งเข้มมาก ยิ่งมีผลต่อการตัดสินใจของแบบจำลอง ส่วนที่สองภาพด้านล่าง มีสีแดงเข้มกระจายที่บริเวณแก้ม ปาก และตา และมีสีฟ้าน้อยมาก ดังนั้นแบบจำลองจึงทำนายว่าภาพนี้เป็นภาพปลอม จากการทำ heatmap ของ SHAP value พบว่า feature สำคัญในภาพ เช่น บริเวณใบหน้า โดยเฉพาะบริเวณดวงตา ริมฝีปาก และแก้ม มักเป็นบริเวณสำคัญที่แบบจำลองให้ความสำคัญในการจำแนกภาพปลอม และภาพจริง โดยเฉพาะในบริบทของภาพที่สร้างโดยปัญญาประดิษฐ์ ซึ่งมักปรากฏลักษณะผิดปกติ หรือความไม่สมมาตร ในบริเวณดังกล่าวจึงส่งผลให้แบบจำลองเรียนรู้และใช้ข้อมูลจากบริเวณเหล่านี้เป็นปัจจัยสำคัญในการตัดสินใจ

ภาพประกอบ 23 ผลลัพธ์ heatmap ของ SHAP value ของแบบจำลอง Swin transformer

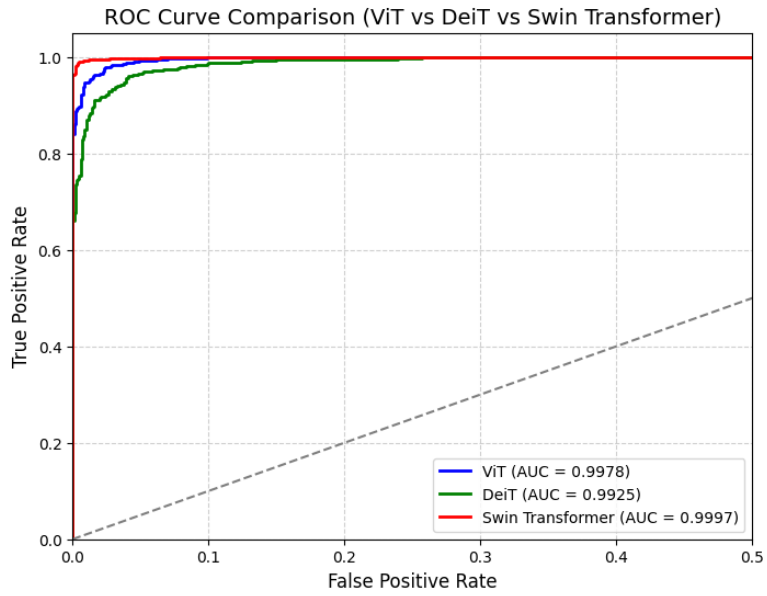
จากภาพประกอบ 24 แสดงภาพตัวอย่างที่แบบจำลอง ViT , DeiT และ Swin transformer ทำนายผิดทั้ง 2 กรณี คือ ภาพจริงทำนายว่าเป็นภาพปลอม และ ภาพปลอมที่ทำนายเป็นภาพจริง ซึ่งความผิดปกติของภาพเป็นปัจจัยสำคัญที่ส่งผลต่อความแม่นยำของแบบจำลองในการจำแนกภาพจริงและภาพปลอม โดยจากงานวิจัยพบว่าคุณภาพของภาพที่ไม่สมบูรณ์อาจทำให้แบบจำลองเกิดข้อผิดพลาดในการทำนาย ปัจจัยหลักที่ส่งผลต่อความผิดพลาดของแบบจำลอง ได้แก่ ความผิดปกติของโทนสี ซึ่งอาจเกิดจากกระบวนการประมวลผลของภาพหรือแหล่งกำเนิดแสงที่ไม่เป็นธรรมชาติ ทำให้แบบจำลองไม่สามารถดึงลักษณะเด่นของใบหน้าได้อย่างถูกต้อง ความมืดของภาพ โดยเฉพาะในบริเวณสำคัญ เช่น ดวงตา ปาก และโครงหน้า ส่งผลให้แบบจำลองไม่สามารถตรวจจับรายละเอียดที่จำเป็นต่อการแยกแยะภาพจริงและภาพปลอมได้อย่างแม่นยำ นอกจากนี้ รายละเอียดใบหน้าที่ไม่ชัดเจน เช่น ภาพเบลอหรือมีแสงเงาซ้อนทับกัน อาจทำให้โครงสร้างใบหน้าไม่สามารถถูกระบุได้อย่างชัดเจน นำไปสู่ข้อผิดพลาดในการทำนาย แหล่งกำเนิดแสงที่ไม่สม่ำเสมอ ก็เป็นอีกปัจจัยที่อาจรบกวนการเรียนรู้ของแบบจำลอง เช่น การมีแสงสะท้อนที่ผิดปกติ หรือความแตกต่างของแสงที่รุนแรงในบางจุดของภาพ เพื่อเพิ่มความแม่นยำของแบบจำลอง อาจจำเป็นต้องมีการปรับปรุงคุณภาพของภาพก่อนป้อนเข้าสู่ระบบ เช่น การปรับสมดุลของแสงและสี หรือการลดสัญญาณรบกวน เพื่อให้แบบจำลองสามารถเรียนรู้ลักษณะที่แท้จริงของภาพได้อย่างมีประสิทธิภาพมากขึ้น



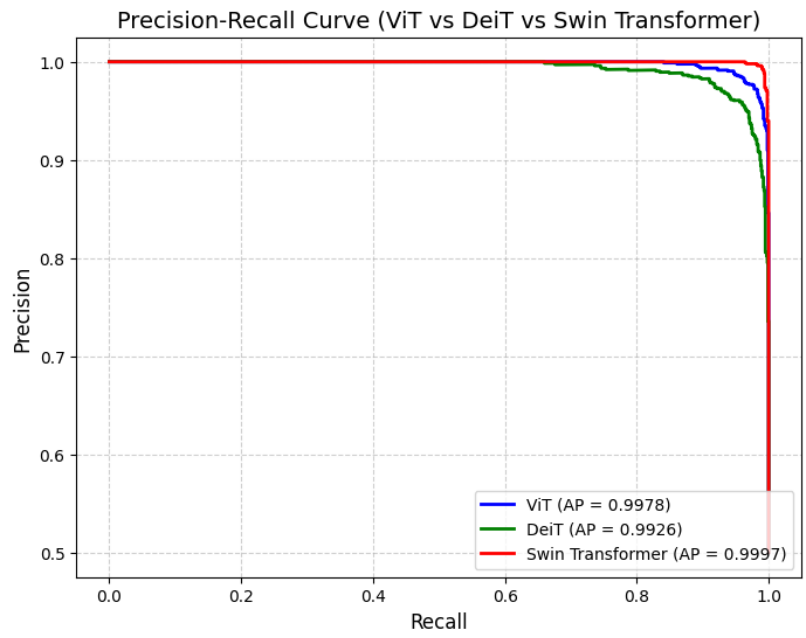
ภาพประกอบ 24 ตัวอย่างภาพที่แบบจำลอง ViT , DeiT และ Swin transformer ทำนายผิด

การเปรียบเทียบค่า AUC Score จาก ROC Curve จาก ภาพประกอบที่ 25 ซึ่งแสดงกราฟ ROC Curve และค่า AUC Score ของแบบจำลอง ViT, DeiT และ Swin Transformer พบว่า Swin Transformer มีค่า AUC Score สูงสุดที่ 0.9997 ซึ่งบ่งชี้ว่าแบบจำลองนี้สามารถจำแนกภาพจริงและภาพปลอมได้อย่างแม่นยำมากที่สุดรองลงมาคือ ViT ที่มีค่า AUC Score 0.9978 ซึ่งแสดงถึงประสิทธิภาพที่ดีในการแยกแยะข้อมูล และสุดท้ายคือ DeiT ที่มีค่า AUC Score 0.9925 แม้ว่าจะต่ำกว่าสองแบบจำลองข้างต้นเล็กน้อย แต่ยังคงอยู่ในระดับที่ดีมาก ค่า AUC Score ที่ใกล้เคียงกับ 1.0 แสดงถึงความสามารถของแบบจำลองในการจำแนกข้อมูลได้อย่างถูกต้อง โดย Swin Transformer มีค่าใกล้เคียง 1 มากที่สุด สะท้อนให้เห็นว่าแบบจำลองนี้สามารถแยกแยะระหว่างภาพจริงและภาพปลอมได้ดีกว่าทั้ง ViT และ DeiT

การเปรียบเทียบค่า AP Score จาก PR Curveจาก ภาพประกอบที่ 26 ซึ่งแสดงกราฟ Precision-Recall (PR) Curve และค่า Average Precision (AP) Score ของแบบจำลอง ViT, DeiT และ Swin Transformer พบว่า Swin Transformer มีค่า AP Score สูงสุดที่ 0.9997 ซึ่งบ่งชี้ว่าแบบจำลองนี้สามารถจำแนกภาพจริงและภาพปลอมได้อย่างแม่นยำมากที่สุดรองลงมาคือ ViT ที่มีค่า AP Score 0.9978 ซึ่งยังคงแสดงถึงความสามารถในการจำแนกข้อมูลที่ดีมาก ส่วน DeiT มีค่า AP Score 0.9926 แม้ว่าจะต่ำกว่าสองแบบจำลองข้างต้น แต่ยังอยู่ในระดับที่สูง ค่า AP Score ที่ใกล้เคียง 1.0 แสดงถึงความสามารถของแบบจำลองในการรักษาความสมดุลระหว่าง Precision และ Recall ได้ดี Swin Transformer มีค่า AP Score สูงที่สุด ซึ่งหมายความว่าแบบจำลองนี้สามารถคงความแม่นยำ (Precision) ได้สูงในขณะที่ยังคงสามารถดึงข้อมูลที่เกี่ยวข้องได้ดี (Recall)



ภาพประกอบ 25 กราฟเปรียบเทียบกราฟ ROC Curve ค่า AUC Score ของโครงข่ายประสาทเทียมทรานฟอร์เมอร์ทั้ง 3 แบบจำลอง (ViT , DeiT และ Swin Transformer)



ภาพประกอบ 26 กราฟเปรียบเทียบกราฟ PR Curve และค่า Average Precision (AP) Score ของโครงข่ายประสาทเทียมทรานฟอร์เมอร์ทั้ง 3 แบบจำลอง (ViT , DeiT และ Swin Transformer

จากผลการทดลองในงานวิจัยนี้หากต้องการนำแบบจำลองที่สร้างไปใช้เพื่อจำแนกภาพจริงและภาพปลอม ผู้วิจัยมีความเห็นว่าแบบจำลองที่ใช้สถาปัตยกรรมแบบ Swin Transformer มีความสามารถในการจำแนกภาพจริงและภาพปลอมได้ดีที่สุด เนื่องจากสามารถเรียนรู้เชิงลึกโดยไม่เกิด Overfitting มากนัก และสามารถจับรายละเอียดของภาพได้ดีมีค่าความแม่นยำ สูงสุดที่ 0.9915 และมีค่า AUC Score และ AP Score ใกล้เคียง 1 มากที่สุด ซึ่งสะท้อนถึงประสิทธิภาพสูงสุดในการจำแนกภาพจริงและภาพปลอม ในขณะที่ ViT และ DeiT มีค่าความแม่นยำรองลงมา ซึ่งแม้ว่าแบบจำลองเหล่านี้จะสามารถจำแนกภาพได้ดี แต่ยังมีข้อจำกัดด้านการเรียนรู้ความสัมพันธ์เชิงพื้นที่ของภาพ และมีแนวโน้มเกิด Overfitting มากกว่า Swin Transformer เนื่องจากต้องใช้ข้อมูลขนาดใหญ่ในการฝึกแบบจำลองเพื่อป้องกันการจดจำข้อมูลมากเกินไป

5.3 ข้อเสนอแนะ

จากผลการวิจัยเกี่ยวกับการใช้แบบจำลองโครงข่ายประสาทเทียมทรานฟอร์เมอร์ทั้ง 3 ประเภท ได้แก่ ViT, DeiT และ Swin Transformer สำหรับการจำแนกภาพจริงและภาพปลอม สามารถสรุปข้อเสนอแนะเพื่อพัฒนางานวิจัยในอนาคตได้ดังนี้

1.การเพิ่มขนาดข้อมูล แม้ว่างานวิจัยมีการทำ Data Augmentation จะช่วยเพิ่มความหลากหลายของข้อมูล แต่การขยายขนาดชุดข้อมูลให้มีความหลากหลายมากยิ่งขึ้นด้วยการเก็บข้อมูลจากแหล่งที่แตกต่าง หรือ ใช้ชุดข้อมูลอื่นที่มีคุณภาพสูงกว่าก็อาจช่วยลดการเกิด Overfittingและ ช่วยเพิ่มประสิทธิภาพของแบบจำลองได้มากยิ่งขึ้น

2. การใช้เทคนิคการ Fine-tuning แม้ว่างานวิจัยมีการทำ Pre-training ที่ช่วยในการฝึกแบบจำลอง แต่การ fine-tune แบบจำลองโดยใช้ข้อมูลเฉพาะที่เกี่ยวข้องกับปัญหาโดยตรง สามารถช่วยปรับปรุงประสิทธิภาพได้มากขึ้น โดยสามารถปรับการเรียนรู้ของแบบจำลองให้มีความแม่นยำสูงสุด

3. การปรับพารามิเตอร์ ควรการปรับแต่ง Hyperparameter เพิ่มเช่นควรศึกษาผลกระทบของ Learning Rate, Batch Size และ Optimizer ต่อผลลัพธ์ของแบบจำลอง เพื่อให้ได้ค่าที่เหมาะสมที่สุดส่งผลให้แบบจำลองทำงานได้มีประสิทธิภาพมากยิ่งขึ้น

4. การปรับปรุงคุณภาพของภาพก่อนการทดสอบ ควรพิจารณาการปรับสมดุลแสงและสี รวมถึงการลดสัญญาณรบกวนก่อนป้อนภาพเข้าสู่ระบบ เพื่อให้แบบจำลองสามารถเรียนรู้ลักษณะที่แท้จริงของภาพได้อย่างมีประสิทธิภาพยิ่งขึ้น

บรรณานุกรม

1. Nguyen TT, Nguyen QVH, Nguyen DT, Nguyen DT, Huynh-The T, Nahavandi S, et al. Deep learning for deepfakes creation and detection: A survey. Computer Vision and Image Understanding. 2022;223:103525.

2. 140k Real and Fake Faces kaggle [Available from: https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces.

3. Team N. Generative AI คืออะไร [Available from: https://tinyurl.com/424utt7n.

4. deepfakes_faceswap. [Available from: https://github.com/deepfakes/faceswap.

5. Nawaz M, Javed A, Irtaza A. A deep learning model for FaceSwap and face-reenactment deepfakes detection. Applied Soft Computing. 2024;162:111854.

6. Mika W. The Emergence of Deepfake Technology: A Review. Technology Innovation Management Review. 2019;9(11).

7. Aldwairi M, Alwahedi A. Detecting Fake News in Social Media Networks. Procedia Computer Science. 2018;141:215-22.

8. FakeApp, 2022. FakeApp 2.2.0. [Available from: https://www.malavida.com/en/soft/fakeapp/.

9. Krichen M, editor Generative Adversarial Networks. 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT); 2023 6-8 July 2023.

10. Karras T, Laine S, Aila T, editors. A Style-Based Generator Architecture for Generative Adversarial Networks. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 15-20 June 2019.

11. Pumarola A, Agudo A, Martinez AM, Sanfeliu A, Moreno-Noguer F. GANimation: Anatomically-aware Facial Animation from a Single Image. Comput Vis ECCV.

2018;11214:835-51.

12. Nirkin Y, Keller Y, Hassner T, editors. FSGAN: Subject Agnostic Face Swapping and Reenactment. 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 27 Oct.-2 Nov. 2019.

13. Zhu JY, Park T, Isola P, Efros AA, editors. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. 2017 IEEE International Conference on Computer Vision (ICCV); 2017 22-29 Oct. 2017.

14. Li Y, Chang MC, Lyu S, editors. In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking. 2018 IEEE International Workshop on Information Forensics and Security (WIFS); 2018 11-13 Dec. 2018.

15. Weerawardana M, Fernando T, editors. Deepfakes Detection Methods: A Literature Survey. 2021 10th International Conference on Information and Automation for Sustainability (ICIAfS); 2021 11-13 Aug. 2021.

16. Güera D, Delp EJ, editors. Deepfake Video Detection Using Recurrent Neural Networks. 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS); 2018 27-30 Nov. 2018.

17. Yuezun L, Siwei L. Exposing DeepFake Videos By Detecting Face Warping Artifacts. 2018.

18. Zhao T, Xu X, Xu M, Ding H, Xiong Y, Xia W, editors. Learning Self-Consistency for Deepfake Detection. 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 10-17 Oct. 2021.

19. Afchar D, Nozick V, Yamagishi J, Echizen I, editors. MesoNet: a Compact Facial Video Forgery Detection Network. 2018 IEEE International Workshop on Information Forensics and Security (WIFS); 2018 11-13 Dec. 2018.

20. Face2Face. [Available from: https://github.com/SocAIty/face2face.

21. Vaswani A, Shazeer NM, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al., editors.

Attention is All you Need. Neural Information Processing Systems; 2017.

22. Ilya S, Oriol V, Quoc VL. Sequence to Sequence Learning with Neural Networks. ArXiv. 2014;abs/1409.3215.

23. Rico S, Barry H, Alexandra B. Neural Machine Translation of Rare Words with Subword Units. ArXiv. 2015;abs/1508.07909.

24. Jonas G, Michael A, David G, Denis Y, Yann D. Convolutional Sequence to Sequence Learning. ArXiv. 2017;abs/1705.03122.

25. Jimmy B, Jamie Ryan K, Geoffrey EH. Layer Normalization. ArXiv. 2016;abs/1607.06450.

26. Transformer ในปัญญาประดิษฐ์คืออะไร [Available from: https://aws.amazon.com/th/what-is/transformers-in-artificial-intelligence/.

27. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ArXiv. 2020;abs/2010.11929.

28. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, J'egou He, editors. Training data-efficient image transformers & distillation through attention. International Conference on Machine Learning; 2020.

29. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al., editors. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 10-17 Oct. 2021.

30. Sharma J, Sharma S, Chahar V, Hussein H, Alshazly H. Deepfakes Classification of Faces Using Convolutional Neural Networks. Traitement du Signal. 2022;39.

31. Ghita B, Kuzminykh I, Usama A, Bakhshi T, Marchang J, editors. Deepfake Image Detection Using Vision Transformer Models. 2024 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom); 2024 24-27 June 2024.

32. Flickr-Faces-HQ Dataset (FFHQ) [Internet]. Available from: https://github.com/NVlabs/ffhq-dataset.

ประวัติผู้เขียน

ชื่อ-สกุล	บุษกร เดชาพงษ์
วัน เดือน ปี เกิด	07 พฤษภาคม 2540
สถานที่เกิด	สุราษฎร์ธานี
วุฒิการศึกษา	พ.ศ.2563 วิศวกรรมศาสตรบัณฑิต สาขาวิชาวิศวกรรมเคมี มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
ที่อยู่ปัจจุบัน	53/61 ม.6 ไอลีพ ไพร์ม พระราม2 กม.14 ซอยแสมดำ ต.พันท้าย อ.เมือง จ.สมุทรสาคร 74000