การทำนายการจดจำภาพด้วยการเรียนรู้เชิงลึก

รัฐพร คุณสมบัติ[1*], นภา แซ่เบ๊[2]

บทคัดย่อ

การจดจำภาพเป็นสิ่งสำคัญในชีวิตประจำวันของมนุษย์ทุกคน เป็นประสบการณ์ที่สร้างความทรงจำและเอกลักษณ์เฉพาะบุคคล โดยการวิจัยมุ่งเน้นทำนายคะแนนการจดจำภาพ โดยนำเสนอเทคนิคการเรียนรู้เชิงลึก(Deep Learning) ได้ฝึกแบบจำลอง 3 แบบ ได้แก่ 1) การฝึกสร้างแบบจำลองจากแรกเริ่ม ศูนย์หรือต้นแบบ (Trained from scratch) 2) การนำฝึกแบบจำลองที่ฝึกมาจากชุดข้อมูลอื่นมาใช้เป็นแบบจำลองเพื่อการคำนวณเวกเตอร์คุณลักษณะของการเรียนรู้เชิงลึกก่อนหน้า (Pretrained model) 3) นำฝึกแบบจำลองที่ฝึกมาจากชุดข้อมูลอื่นเพื่อมาการปรับแต่งเพิ่มเติมแบบจำลองที่ได้ฝึกไว้ก่อนหน้า (Fine-tuning) ซึ่งใช้โครงสร้างจากโครงข่ายแบบสังวัฒนาการ (Convolutional Neural Network, CNN) โดยสกัดคุณลักษณะ (Feature Extraction) จากแบบจำลอง ResNet 50 และ ใช้โครงข่ายแบบ Transformer โดยสกัดคุณลักษณะ จากแบบจำลอง Vision Transformer และทำการนำเวกเตอร์คุณลักษณะทั้งสองแบบจำลองมาเชื่อมต่อกัน ซึ่งดำเนินการวิจัยโดยใช้ชุดข้อมูล 2 รูปแบบ ได้แก่ แบบเฉพาะหมวดหมู่และแบบคละหมวดหมู่ เพื่อทำการเปรียบเทียบจากแบบจำลองทั้งสามแบบ โดยนำเวกเตอร์คุณลักษณะที่ได้มาใช้ในการทำนายการจดจำภาพแบบถดถอย (Regression) ผลลัพธ์จากการทดลองเปรียบเทียบผลการทำนายระหว่างแบบจำลองดังกล่าว ทำงานได้ดีกับชุดข้อมูลที่มีการฝึกแบบแบบคละหมวดหมู่ ซึ่งแบบจำลองที่ได้ผลออกมาที่ดีสุด คือ แบบจำลอง ResNet 50 โดยการฝึกแบบจำลองที่ฝึกเพิ่มเติมจากแบบจำลองที่ฝึกมาแล้ว มีค่าประสิทธิภาพการทำนายคะแนนการจดจำภาพดังนี้ Mean Squared Error (MSE) 0.001 Mean Absolute Error (MAE) 0.0082 R-square (R2) 0.9947 และ Spearman Correlation Coefficient (Spearman's rho) 0.9896

คำสำคัญ : คะแนนการจดจำภาพ , การทำนายการจดจำภาพ, เทคนิคการเรียนรู้เชิงลึก

---

[1] หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ กรุงเทพฯ 10110

[2] คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ กรุงเทพฯ 10110

* Corresponding author:Tel.: 065-5070567 E-mail address: rattaporn.kun@g.swu.ac.th

Image Memorability Prediction using Deep Learning

Rattaporn Kunsombat[1*], Napa Sae-bae[2]

Abstract

The image memorability plays a significant role in everyday human life, as it generates memorable experiences and individual identities. Research in this domain focuses on predicting image memorability scores using deep learning techniques. Three model architectures were trained and evaluated: 1) models trained from scratch, 2) models pretrained on anther dataset, and 3) fine-tuned models based on pretrained architectures. These models utilized structures from Convolutional Neural Networks (CNNs) for feature extraction, with features extracted from ResNet 50 and Vision Transformer models. The extracted feature vectors from both models were then concatenated. The research was conducted using two types of datasets: category-specific and mixed-category datasets, allowing for a comparative analysis of the three model types. The experiments demonstrated superior performance on mixed-category datasets, with the best-performing model being the fine-tuned ResNet 50 model. This model achieved the following performance metrics for predicting image memorability : Mean Squared Error (MSE) of 0.001, Mean Absolute Error (MAE) of 0.0082, R-square (R2) of 0.9947, and Spearman Correlation Coefficient (Spearman's rho) of 0.9896.

Keywords: image recognition score, image recognition prediction, deep learning techniques

[1] Data Science, Faculty of Science, Srinakharinwirot University, Bangkok, 10110, Thailand

[2] Faculty of Science, Srinakharinwirot University, Bangkok, 10110, Thailand

* Corresponding author:Tel.: 065-5070567 E-mail address: rattaporn.kun@g.swu.ac.th

บทนำ

การจดจำภาพเป็นสิ่งสำคัญในชีวิตประจำวันของมนุษย์ทุกคนซึ่งประสบการณ์ที่เป็นเอกลักษณ์และสร้างความทรงจำเป็นของตนเองมักเกิดขึ้นเมื่อพบเจอสิ่งต่าง ๆ บางสิ่งอาจจดจำได้ทันที ในขณะที่บางสิ่งก็อาจลืมได้ในระยะเวลาอันสั้น ความสามารถในการจดจำภาพสามารถวัดได้จากพฤติกรรมและประสบการณ์ของแต่ละบุคคล โดยมุมมองทางจิตวิทยาความจำมาจากการกระตุ้นภายในสมองวิทยาและการใช้ชีวิตประจำวัน ในความเป็นจริง มนุษย์ไม่สามารถที่จะคาดการณ์ได้ว่าจะสามารถจดจำสิ่งต่าง ๆ ได้ดีหรือไม่ งานวิจัยอื่น [1] พบว่าในแต่ละบุคคลมักจะมีความสอดคล้องกันในภาพที่สามารถจดจำได้ นั่นคือ ผู้คนมักจะจดจำและลืมภาพแบบเดียวกัน แม้ว่าจะมีประสบการณ์ที่แตกต่างกัน เช่น ใบหน้าและฉาก เราสามารถสร้างแนวคิดเกี่ยวกับการจดจำโดยวัดผลจากความน่าจะเป็นที่แต่ละบุคคลจะจำภาพนั้นได้หลังจากที่พบเห็นหรือไม่ ดังนั้น แม้ว่ามนุษย์จะไม่สามารถคาดการณ์การจดจำของภาพได้ แต่ในงานวิจัยนั้น [2] นักวิจัยสามารถให้คำตอบว่ามนุษย์จะจดจำภาพใดได้บ้าง โดยพิจารณาจากความสามารถในการจดจำที่วัดได้เพียงอย่างเดียว ซึ่งความทรงจำในความเป็นจริงนั้น การรู้ว่าภาพใดที่น่าจะถูกจดจำโดยไม่คำนึงถึงลักษณะของผู้สังเกตการณ์สามารถเป็นประโยชน์และต่อยอดในการทำงานอื่น ๆ ได้ ในทางหนึ่ง เราสามารถใช้ความรู้นี้เพื่อจัดการกับความทรงจำ มีการใช้งานด้านการจดจำภาพนี้ในงานสาขาต่าง ๆ มากมาย เช่น การศึกษา การโฆษณาและสื่อ การแพทย์

จากเหตุผลที่กล่าวมาข้างต้น ผู้วิจัยเล็งถึงความสำคัญของการจดจำภาพในชีวิตประจำวันของมนุษย์ การจดจำนี้สามารถมีผลประโยชน์ได้ในหลายมิติต่าง ๆ ดังนั้น ผู้วิจัยได้ทำการสร้างแบบจำลองทำนายการจดจำภาพโดยใช้เทคนิคการเรียนรู้เชิงลึกสามรูปแบบ โดยรูปแบบที่หนึ่งอาศัยโครงข่ายประสาทเทียมแบบสังวัตนาการ ที่เรียกว่า ResNet 50 รูปแบบที่สองอาศัยโครงสร้างแบบจำลอง Transformer ที่เรียกว่า Vision Transformer มาใช้ในการประมวลผลเพื่อคำนวณเวกเตอร์คุณลักษณะที่นำไปใช้ในการทำนายการจดจำและรูปแบบที่สามอาศัยการนำเวกเตอร์คุณลักษณะทั้งสองโมเดลมาเชื่อมต่อกันและทำการเปรียบเทียบผลการทำนายของแบบจำลองทั้งสามรูปแบบโดยนำเวกเตอร์คุณลักษณะจากแบบจำลองทั้งสามประเภทมาการทำนายการจดจำภาพแบบถดถอย

งานวิจัยที่เกี่ยวข้อง

บทความวิจัยที่เกี่ยวข้องกับการจดจำภาพโดยใช้เทคนิคการเรียนรู้เชิงลึก มีตัวอย่างดังต่อไปนี้ [3] บทความวิจัยนี้กล่าวถึง ชุดข้อมูล “LaMeM” ซึ่งเป็นชุดข้อมูลความจำภาพที่มีคำอธิบายประกอบเชิงความหมายที่ใหญ่ที่สุด ประกอบด้วย 60,000 ภาพจากแหล่งที่มาที่หลากหลายโดยใช้ โครงข่ายประสาทเทียม Convolutional Neural Networks (CNN) และคุณสมบัติเชิงลึกที่ปรับแต่งอย่างละเอียดเพื่อประเมินความสามารถในการจดจำของภาพ มีประสิทธิภาพเหนือกว่าคุณสมบัติอื่น ๆ และอันดับ Spearman เท่ากับ 0.64 ซึ่งมีความสอดคล้องใกล้เคียงกับมนุษย์

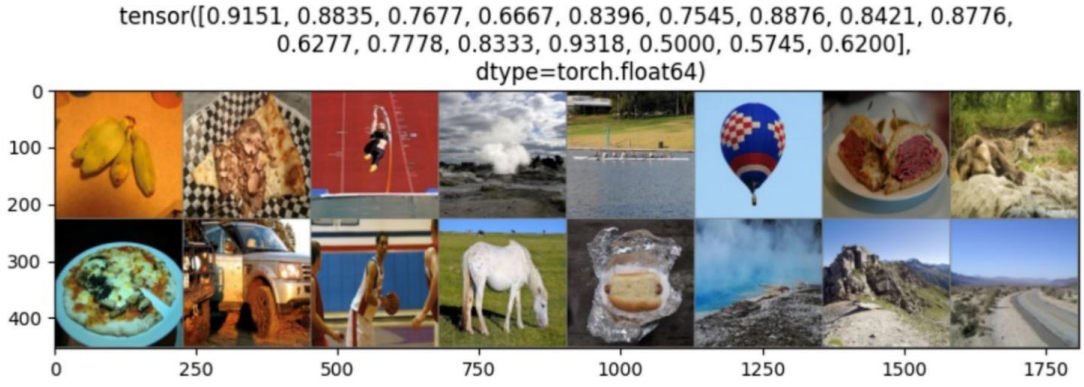
การวิจัยแสดงให้เห็นถึงการประมาณความสามารถในการจดจำภาพที่ดีในชั้นต่างๆ ความสามารถในการจดจำตำแหน่งและคุณสมบัติของภาพ มีการปรับรายละเอียดในส่วนต่างๆของแบบจำลอง เพื่อประเมินความสามารถในการจดจำของภาพซึ่งมีประสิทธิภาพดีกว่า ในแบบจำลองอื่นๆ

การวิจัยอาศัยคำอธิบายประกอบเชิงความหมายเป็นหลักเพื่อสร้างชุดข้อมูล LaMem ซึ่งอาจมีส่วนในความสามารถในการจดจำในแต่ละบุคคลได้ โดยการวิจัยไม่ได้สำรวจผลกระทบของปัจจัยบริบทเช่นคำบรรยายภาพหรือข้อความโดยรอบต่อความสามารถในการจดจำภาพ [4] บทความนี้มุ่งเน้นการศึกษาและวิเคราะห์ความจำของ วัตถุในภาพ และสำรวจความสัมพันธ์ระหว่างความจำของวัตถุและภาพ งานวิจัยนี้ได้รวบรวมข้อมูลจริงเพื่อเข้าใจปัจจัยที่มีผลต่อความจำของวัตถุ เช่น ประเภทของวัตถุและความชัดของภาพและยังสำรวจความสัมพันธ์ของการจดจำระหว่างภาพกับวัตถุ โดยใช้แบบจำลองการเรียนรู้เชิงลึก Conv-net ที่ได้รับการฝึกบนชุดข้อมูล ImageNet ถูกนำมาใช้เพื่อทำนายความสามารถในการจดจำวัตถุในภาพ ซึ่งคุณสมบัติรูปภาพ ยังใช้ SIFT และ HOG เพื่อทำนายความสามารถในการจดจำวัตถุ [5] บทความนี้เสนอสถาปัตยกรรมการเรียนรู้เชิงลึกแบบใหม่ที่เรียกว่า ResMem-Net ซึ่งเป็นผสมผสานระหว่าง LSTM และ CNN โดยแบบจำลอง ResMem-Net มาจากโครงสร้างพื้นฐานของ ResNet (Residual Network) ซึ่งแบบจำลองนี้ได้รับการฝึกอบรมและประเมินผลโดยใช้ชุดข้อมูล Large-scale Image Memorability (LaMem) ผลลัพธ์พบว่า ResMem-Net มีความสัมพันธ์อันดับ 0.679 และค่าคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) 0.011 ซึ่งมีประสิทธิภาพที่ดี [6] บทความวิจัย ได้มีการสำรวจการประยุกต์ใช้โครงสร้างแบบจำลอง Transformer ที่ใช้กันอย่างแพร่หลายในการประมวลผลภาษาธรรมชาติ มีการใช้งานที่จำกัดในการประมวลผลรูปภาพ โดยวิธีการเดิมในการประมวลผลรูปภาพนั้น มักจะรวม attention กับ convolutional networks หรือแทนบางส่วนของ convolutional networks ด้วย attention งานวิจัยนี้ได้นำแบบจำลอง transformer ที่ใช้งานโดยตรงกับลำดับของ image patches สำหรับงานการจำแนกภาพ เรียกแบบจำลองว่า Vision Transformer (ViT) ที่ทำการฝึกรูปแบบ pre-trained ด้วยข้อมูลจำนวนมากและโอนย้ายไปยัง benchmarks ซึ่งการจำแนกภาพต่างๆ ได้ผลลัพธ์เป็นอย่างดี เมื่อเปรียบเทียบกับ convolutional networks ที่เป็น state-of-the-art และยังใช้ทรัพยากรคำนวณน้อยกว่าอีกด้วย

วิธีดำเนินการ

ขั้นตอนที่ 1 : ชุดข้อมูลที่ใช้ในการศึกษานี้

ผู้วิจัยนำชุดข้อมูล “ MemCat” (Lore Goetschalckx and Johan Wagemans ,2019) ประกอบด้วยชุดข้อมูลรูปภาพจำนวน 10,000 รูป แบ่งเป็น 5 หมวดหมู่ หมวดละ 2,000 รูป ได้แก่ Animal, Food, Sports, Landscape, Vehicle และชุดข้อมูลในรูปแบบตารางที่ระบุรายละเอียดเกี่ยวกับชื่อรูปภาพ,หมวดหมู่,หมวดหมู่ย่อย,ความ,กว้างและสูงของรูป,แหล่งที่มาของรูปภาพ,ป้ายกำกับภาพ จากแหล่งที่มา,จำนวนการเข้าดูรูป,การแจ้งเตือนที่ผิดพลาดจากการจดจำรูปภาพ,จำนวนผู้เข้าร่วมการจดจำภาพที่ผ่านเกณฑ์,คะแนนความจดจำภาพ โดยในการทดลองนี้จะใช้ข้อมูลเฉพาะ ชื่อรูปภาพ ,หมวดหมู่ และคะแนนความจดจำภาพ



ภาพประกอบ 1 ตัวอย่างภาพที่ใช้ในการฝึกแบบจำลอง

ขั้นตอนที่ 2 : การเตรียมรูปภาพ

กระบวนการปรับรูปภาพให้เหมาะสมก่อนนำเข้าแบบจำลองในการทดลองได้แก่การปรับขนาดภาพให้เป็น 224*224 พิกเซล โดยเริ่มจากเติมขอบดำให้ภาพ (Zero Padding) สำหรับภาพต้นฉบับที่ไม่เป็นจัตุรัส ก่อนการปรับขนาดภาพ เพื่อให้ภาพที่นำเข้ามีความสูงของภาพเท่ากันกับความกว้างที่ 224*224 พิกเซลเพื่อให้เข้ากันได้กับแบบจำลองที่นำมาใช้โดยไม่เป็นการกระทบกับอัตราส่วนของภาพ

ขั้นตอนที่ 3 : การสร้างแบบจำลอง

งานวิจัยนี้ผู้วิจัยใช้เทคนิคการเรียนรู้เชิงลึก 3 แบบ คือ แบบจำลอง ResNet50, Vision Transformer, และการเชื่อมต่อกันระหว่างสองแบบจำลอง โดยได้เลือกใช้ไลบรารีของ PyTorch ในการสร้างแบบจำลองเนื่องจากมีความยืดหยุ่นในการสร้างแบบจำลองสำหรับการเรียนรู้เชิงลึกโดยได้ทำการฝึกแบบจำลองทั้งหมด 3 แบบ ดังนี้

1. การฝึกแบบจำลองจากแรกเริ่ม (Trained from scratch)
2. การนำแบบจำลองที่ฝึกมาจากชุดข้อมูลอื่นมาใช้เป็นแบบจำลองเพื่อการคำนวณเวกเตอร์คุณลักษณะ (Pretrained model)
3. การนำฝึกแบบจำลองที่ฝึกมาจากชุดข้อมูลอื่นมาปรับแต่งเพิ่มเติม (Fine-tuning)

ซึ่งผู้วิจัยดำเนินการสกัดคุณลักษณะ โดยใช้แบบจำลอง ResNet 50 จากแบบจำลองโครงข่ายแบบสังวัฒนาการซึ่งใช้การกรอง (Filter) เพื่อสกัดคุณลักษณะจากภาพ และแบบจำลอง Vision Transformer จากแบบจำลอง Transformer โดยใช้ทำการแบ่งรูปภาพออกเป็น patch เล็กๆ และแปลงเป็น embedding patch หลังจากนั้นนำเข้า encoder layers เพื่อทำการประมวลผลและนำไปสู่ layer ในชั้นอื่นๆ โดยในการสกัดคุณลักษณะทั้งสองแบบดังกล่าวจะได้ layer ในชั้นสุดท้าย (output) เป็น Regression โดยมีการกำหนดพารามิเตอร์ ดังตารางที่ 1

ตารางที่ 1 พารามิเตอร์ที่ใช้ในการสร้างแบบจำลอง

Parameter	Value
Batch size	32
Activation	Sigmoid
Optimizer	Adam

โดยจะมีการแบ่งข้อมูลเพื่อทำการทดลองด้วยอัตราส่วน Train set 90% Validation Set 5% และ Test Set 5% ของรูปภาพ และจากชุดข้อมูลที่แยกตามหมวดหมู่ 5 หมวดหมู่ หมวดหมู่ละ 2,000 รูป และจากชุดข้อมูลแบบรวมทั้ง 5 หมวดหมู่ทั้งหมด 10,000 รูป

ขั้นตอนที่ 4 : การประเมินผลแบบจำลอง

วัดประสิทธิภาพโดยใช้ข้อมูลที่แบบจำลองทำนายและค่าเป้าหมายจากชุดข้อมูลทดสอบซึ่งมี 4 ตัวชี้วัด ดังนี้

1. Mean Squared Error (MSE) เป็นการวัดว่าราคาที่ทำนายนั้นไกลจากราคาจริงแค่ไหน ซึ่งคำนวณโดยการหาค่าเฉลี่ยของความแตกต่างระหว่างราคาที่ทำนายกับราคาจริงในรูปกำลังสอง

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Yi - \widehat{Yi})^2$$

จากสมการ

Yi คือ ค่าจริง (actual value) ของ samples test ที่ i

$\widehat{Yi}$ คือ ค่าที่ประมาณได้หรือ ค่า predict ของ samples test ที่ i

n คือ จำนวน samples ทั้งหมด

2. Mean Absolute Error (MAE) หรือ ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean Absolute Error: MAE) คือ การหาค่าเฉลี่ยของความแตกต่างสมบูรณ์ระหว่างค่าทำนายและค่าจริง หากค่า MAE นั้นมีค่าน้อย แสดงว่าค่าทำนายนั้นมีค่าใกล้เคียงกับค่าจริง ดังสมการ

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|yi - \hat{y}|$$

จากสมการ

Yi คือ ค่าจริง (actual value) ของ samples test ที่ i

$\widehat{Yi}$ คือ ค่าที่ประมาณได้หรือ ค่า predict ของ samples test ที่ i

n คือ จำนวน samples ทั้งหมด

3. R-squared (R^2) เป็นการวัดว่าแบบจำลองอธิบายความแปรปรวนของราคาจริงได้ดีหรือไม่ ซึ่งคำนวณโดยการเปรียบเทียบความแปรปรวนของราคาที่ทำนายกับความแปรปรวนของราคาจริง
โดยที่ ค่าเข้าใกล้ 0 หมายถึง แบบจำลองไม่สามารถอธิบายข้อมูลใดๆ ทั้งสิ้น
ค่าเข้าใกล้ 1 หมายถึง แบบจำลองสามารถอธิบายข้อมูลได้เพียงพอทุกประการ

$$R^2 = 1 - \frac{SSres}{SStot}$$

4. Spearman Correlation Coefficient (Spearman's rho) ค่าที่เกี่ยวข้องกับการคำนวณความสัมพันธ์ระหว่างตัวแปรโดยใช้ค่าความสัมพันธ์ที่ดีของสปีร์แมน (Spearman correlation coefficient) ซึ่งบ่งชี้ถึงความสัมพันธ์ระหว่างลำดับของข้อมูล โดยทั่วไปแล้วค่าของ Spearman correlation coefficient จะอยู่ในช่วง -1 ถึง 1 โดยค่าบวกแสดงถึงความสัมพันธ์ที่เชิงบวก ค่าลบแสดงถึงความสัมพันธ์ที่เชิงลบ และค่าเป็นศูนย์แสดงถึงขาดความสัมพันธ์ การใช้ Spearman correlation coefficient เป็นที่นิยมในการวิเคราะห์ข้อมูลที่มีการจัดลำดับ อาทิ การวิเคราะห์ข้อมูลที่มีการจัดอันดับของคะแนนหรือการจัดอันดับของผลการทดลอง โดยมีสมการดังนี้

$$\rho = 1 - \frac{6\Sigma\, \mathrm{d}_i^2}{n(n^2-1)}$$

จากสมการ

ρ คือ Spearman correlation coefficient

d_i^2 คือ ความต่างระหว่างความลำบากที่ปรากฏในอันดับของข้อมูลที่เปรียบเทียบกันระหว่างคู่ของตัวแปร

n คือ จำนวนข้อมูล

ผลการวิจัยและอภิปรายผลการวิจัย

1. การฝึกแบบจำลองจากแรกเริ่มจากการฝึกแบบจำลองใช้เวลาในการฝึกเป็นเวลานานกว่าการฝึกรูปแบบอื่นๆ จึงฝึกแบบจำลองรูปแบบรวมหมวดหมู่โดยใช้ชุดข้อมูลทดสอบทดลองในแบบจำลอง ResNet50, Vision Transformer, และ Vision Transformer + ResNet50 มีค่าดังตารางที่ 2 ซึ่งแบบจำลองจากแรกเริ่มนั้นไม่สามารถอธิบายข้อมูลได้อย่างเพียงพอ ซึ่งยังมีความจำเป็นในการปรับปรุงแบบจำลองเพิ่มเติมเพื่อให้สามารถอธิบายข้อมูลได้อย่างเหมาะสมยิ่งขึ้นในการฝึกรูปแบบอื่นๆ

ตารางที่ 2 แสดงผลการทดลอง(Scratch)

Category	Model	MSE	R-square	MAE
Merge	Vision Transformer	1.1983	-69.023	1.084
	ResNet50	0.353	-19.5287	0.5707
	Vision Transformer + ResNet50	0.7155	-40.88	0.801

2. การนำแบบจำลองที่ฝึกมาจากชุดข้อมูลอื่นมาใช้เป็นแบบจำลองเพื่อการคำนวณเวกเตอร์คุณลักษณะทำการฝึกกับชุดข้อมูลทั้งหมด 5 หมวดหมู่ และนำผลลัพธ์ที่ได้มาคำนวณเพื่อหาค่าเฉลี่ยของแบบจำลองเพื่อเปรียบเทียบกับชุดข้อมูลแบบ Merge all categories โดยใช้ชุดข้อมูลทดสอบในการวัดผล จากแบบจำลอง ResNet50, Vision Transformer, และ Vision Transformer + ResNet 50 ดังตารางที่3

ตาราง 3 แสดงผลการทดลอง (Pre-trained) ของของแบบจำลองทั้ง 3 แบบ

Category	MSE	R-square	MAE	Spearman's
Vision Transformer				
Average 5 Categories	0.0003	0.8487	0.0159	0.6422
	(0.0012)	(0.9132)	(0.0216)	
Merge All Categories	**0.0003**	**0.9870**	**0.0076**	**0.9716**
	(0.0037)	**(0.8310)**	**(0.0500)**	
ResNet 50				
Average 5 Categories	0.0003	0.9522	0.0156	0.8162
	(0.0002)	(0.9301)	(0.0114)	
Merge All Categories	**0.0002**	**0.9945**	**0.0065**	**0.9839**
	(0.0001)	**(0.9919)**	**(0.0098)**	
Vision Transformer + ResNet50				
Average 5 Categories	0.0002	0.95264	0.0091	0.9088
	(0.0008)	(0.9080)	(0.0235)	
Merge All Categories	**0.0001**	**0.936**	**0.0101**	**0.9761**
	(0.0005)	**(0.8570)**	**(0.0171)**	

( ) เป็นการรายงานประสิทธิภาพของแบบจำลองบนชุดข้อมูลสำหรับฝึก

3. การนำฝึกแบบจำลองที่ฝึกมาจากชุดข้อมูลอื่น,มาปรับแต่งเพิ่มเติมทำการฝึกกับชุดข้อมูลทั้งหมด 5 หมวดหมู่ และนำผลลัพธ์ที่ได้มาคำนวณเพื่อหาค่าเฉลี่ยของแบบจำลองเพื่อเปรียบเทียบกับชุดข้อมูลแบบ Merge all categories โดยใช้ชุดข้อมูลทดสอบในการวัดผล จากแบบจำลอง ResNet50, Vision Transformer, และการเชื่อมต่อกันระหว่างสองแบบจำลอง ดังตารางที่ 4

ตาราง 4 แสดงผลการทดลอง (Fine-Tuning model) ของของแบบจำลองทั้ง 3 แบบ

Category	MSE	R-square	MAE	Spearman's
Vision Transformer				
Average 5 Categories	0.0003	0.92904	0.00968	0.8519
	(0.0004)	(0.9457)	(0.0138)	
Merge All Categories	**0.0006**	**0.9864**	**0.0082**	**0.9517**
	(0.0001)	**(0.9930)**	**(0.0093)**	
ResNet 50				
Average 5 Categories	0.0002	0.9619	0.0094	0.8953
	(0.0004)	(0.9020)	(0.0168)	
Merge All Categories	**0.0001**	**0.9947**	**0.0082**	**0.9896**
	(0.0001)	**(0.9916)**	**(0.0097)**	
Vision Transformer + ResNet50				
Average 5 Categories	0.0002	0.9419	0.0123	0.8163
	(0.0005)	(0.9270)	(0.0170)	
Merge All Categories	**0.0002**	**0.9279**	**0.0105**	**0.9523**
	(0.0003)	**(0.9332)**	**(0.0143)**	

( ) เป็นการรายงานประสิทธิภาพของแบบจำลองบนชุดข้อมูลสำหรับฝึก

จากตารางที่ 3 และ 4 แสดงเปรียบเทียบค่า MSE , R-square และMAE และ Spearman's ระหว่างโมเดล Vision Transformer, ResNet50, และVision Transformer + ResNet 50 ได้แสดงให้เห็นว่าแบบจำลองทั้ง 3 แบบทำงานได้ดีในชุดข้อมูลที่มีการรวมกันของหมวดหมู่ทั้ง 5 หมวด ประสิทธิภาพของแบบจำลองที่ใช้ในการทำนายคะแนนการจดจำนั้น พบว่ามีความใกล้เคียงกันระหว่างแบบจำลอง ซึ่งแบบจำลองที่มีประสิทธิภาพสูงสุดนั่นคือ ResNet 50 โดยการฝึกในรูปแบบ Pretrained เนื่องจากเปรียบเทียบในภาพรวมจากการฝึกแบบจำลองอาทิ ค่า MAE ของแบบจำลอง ResNet50 จากการ Fine-tuning ดีกว่า Pretrained แต่เมื่อนำมาทดสอบกลับมีประสิทธิภาพที่แย่ลง สาเหตุอาจเกิดจากการ Fine-tuning ที่ทำให้เกิดการปัญหา overfitting

สรุปผลการวิจัย

แบบจำลอง ResNet50 โดยใช้ Pretrained มีประสิทธิภาพดีที่สุดเมื่อเทียบกับแบบจำลองอื่นๆ ซึ่งการสร้างแบบจำลอง Scratch มีประสิทธิภาพระดับต่ำที่สุด และสามารถสรุปผลได้ดังนี้

1. การฝึกโดยนำแบบจำลองที่ฝึกมาจากชุดข้อมูลอื่นมาใช้เป็นแบบจำลองเพื่อการคำนวณเวกเตอร์คุณลักษณะ มีประสิทธิภาพดีกว่า การฝึกแบบจำลองที่ฝึกมาจากชุดข้อมูลอื่นมาปรับแต่งเพิ่มเติม และการฝึกแบบจำลองจากแรกเริ่ม
2. ประสิทธิภาพของแบบจำลองการวัดผลออกมาในทิศทางเดียวกันทั้งในการฝึกมาจากชุดข้อมูลอื่นมาใช้เป็นแบบจำลองเพื่อการคำนวณเวกเตอร์คุณลักษณะ และการฝึกแบบจำลองที่ฝึกมาจากชุดข้อมูลอื่นมาปรับแต่งเพิ่มเติม คือชุดข้อมูลแบบรวมหมวดหมู่ทำงานได้ดีกว่า
3. จากสมมติฐานว่าการจดจำภาพได้ในแต่ละบุคคลมีความสอดคล้องกัน ในแต่ละบุคคลโดยการจดจำภาพนั้น แบบจำลอง Vision Transformer สามารถทำนายการจดจำได้ดีเทียบเท่า แบบจำลองแบบสังวัฒนาการทั้งนี้การเลือกใช้ความเหมาะสมของข้อมูล ชนิดของข้อมูล จำนวนของข้อมูล อีกทั้งจากการวิจัยพบว่า การจดจำภาพนั้นเป็นส่วนหนึ่งของประสบการณ์ที่แต่ละบุคคลพบเจอของวัตถุหรือ สามารถจดจำองค์ประกอบที่อยู่ภายในภาพได้

กิตติกรรมประกาศ

การจัดทำวิจัยได้รับการสนับสนุนจากบัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ ในการนำเสนอผลงานวิจัย ผู้วิจัยจึงขอขอบคุณมา ณ ที่นี้

เอกสารอ้างอิง

[1] Hagen, T., & Espeseth, T. (2023). Image Memorability Prediction with Vision Transformers. ArXiv, abs/2301.08647.

[2] P. Isola, J. Xiao, D. Parikh, A. Torralba and A. Oliva, What Makes a Photograph Memorable?,.IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 7, pp. 1469-1482, July 2014, doi: 10.1109/TPAMI.2013.200.

[3] A. Khosla, A. S. Raju, A. Torralba and A. Oliva, "Understanding and Predicting Image Memorability at a Large Scale," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 2390-2398, doi: 10.1109/ICCV.2015.275.

[4] R. Dubey, J. Peterson, A. Khosla, M. -H. Yang and B. Ghanem, "What Makes an Object Memorable?," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 1089-1097, doi: 10.1109/ICCV.2015.130.

[5] Praveen A, Noorwali A, Samiayya D, Zubair Khan M, Vincent P M DR, Bashir AK, Alagupandi V. 2021. ResMem-Net: memory based deep CNN for image memorability estimation. PeerJ Computer Science 7:e767 https://doi.org/10.7717/peerj-cs.767

[6] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ArXiv, abs/2010.11929.