การวิเคราะห์ความคงทนของตัวแบบการเรียนรู้เชิงลึก ต่อการโจมตีแบบพอยซันนิ่งแบบแกนส์ ในงานภาพทางการแพทย์

นายภาคภูมิ สิงขรภูมิ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติ ภาควิชาสถิติ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2565

An Analysis of Deep Learning Model's Robustness against GANS-based Poisoning Attack in Medical Imaging

Mr. Pakpoom Singkorapoom

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Statistics
Department of Statistics
FACULTY OF COMMERCE AND ACCOUNTANCY
Chulalongkorn University
Academic Year 2022

หัวข้อวิทยานิพนธ์	การวิเคราะห์ความคงทนของตัวแบบการเรียนรู้เชิงลึกต่อการโจมตีแบบพอยซันนิ่งแบบแกนส์ในงานภาพทางการแพทย์
โดย	นายภาคภูมิ สิงขรภูมิ
สาขาวิชา	สถิติ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร.สุรณพีร์ ภูมิวุฒิสาร

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

..... คณบดีคณะพาณิชยศาสตร์และการบัญชี

(ศาสตราจารย์ ดร.วิเลิศ ภูริวัชร)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ

(รองศาสตราจารย์ ดร.วรธนน์ วิริยสิทธาวัฒน์)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(ผู้ช่วยศาสตราจารย์ ดร.สุรณพีร์ ภูมิวุฒิสาร)

..... กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.ภูริพันธุ์ รุจิขจร)

..... กรรมการภายนอกมหาวิทยาลัย

(ดร.ชาลี ธรรมรัตน์)

ภาคภูมิ สิงขรภูมิ : การวิเคราะห์ความคงทนของตัวแบบการเรียนรู้เชิงลึกต่อการโจมตีแบบพอยซันนิ่งแบบแกนส์ในงานภาพทางการแพทย์.
(An Analysis of Deep Learning Model's Robustness against GANS-based Poisoning Attack in Medical Imaging) อ.ที่ปรึกษาหลัก : ผศ. ดร.สุรณพีร์ ภูมิวุฒิสาร

ปัจจุบันเทคโนโลยี deep learning ได้เข้ามีส่วนช่วยในการพัฒนางานทางด้านการแพทย์และสาธารณสุขเป็นอย่างมาก ด้วยการใช้สถาปัตยกรรมที่ล้ำสมัยและพารามิเตอร์ที่ถูกสอนด้วยข้อมูลขนาดใหญ่ แต่ทว่า model เหล่านี้สามารถถูกโจมตีได้ด้วย adversarial attack เพราะว่า model เหล่านี้ยังต้องพึ่งพารามิเตอร์ในการสร้างเอาต์พุตและลักษณะที่ไม่สามารถอธิบายได้ของ model นั้นก็ทำให้ยากที่จะหาทางแก้หากถูกโจมตีแล้ว ในทุกๆวันมีการใช้ model เหล่านี้เยอะมากขึ้นเพื่อช่วยบุคลากรทางการแพทย์ แต่ด้วยงานที่ต้องคำนึงถึงชีวิตของผู้คนเป็นหลักการทดสอบความปลอดภัยและความคงทนของตัว model จึงจำเป็น การโจมตีสามารถแบ่งได้ออกเป็นสองประเภทคือ evasion atttack และ poisoning attack ที่มีความยืดหยุ่นกว่า evasion attack ทั้งในเรื่องของการสร้างข้อมูลแปลกปลอมใหม่ขึ้นมาและวิธีการโจมตีทำให้การทดสอบความคงทนต่อ poisoning attack ในงานทางการแพทย์นั้นสำคัญเป็นอย่างยิ่ง วิทยานิพนธ์ฉบับนี้ศึกษาความคงทนของ deep learning model ที่มีสถาปัตยกรรมล้ำสมัยที่ถูกพัฒนามาเพื่องานจำแนกภาพเอกซเรย์ปอดแบบไบนารีภายใต้การโจมตีแบบ poisoninng attack การโจมตีนั้นจะใช้ GANs ในการสร้างข้อมูลสังเคราะห์ปลอมขึ้นมาและทำการติดป้ายกำกับที่ผิดให้ในรูปแบบของ black box และใช้ปริมาณของตัววัดที่ลดลงเมื่อนำข้อมูลนี้ไปอัพเดท model เป็นตัวบ่งชี้ถึงคความคงทนของแต่ละสถาปัตยกรรมที่แตกต่างกันออกไป จากการทดลองเราพบว่าสถาปัตยกรรม ConvNext นั้นมีความคงทนมากที่สุดและอาจจะสื่อได้ว่าเทคโนโลยีที่มาจาก Transformer นั้นมีส่วนช่วยสนับสนุนความคงทนของ model

สาขาวิชา สถิติ ลายมือชื่อนิสิต .....
ปีการศึกษา 2565 ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

6380259226 : MAJOR STATISTICS
KEYWORD: adversarial machine learning poisoning attack medical image classification generative adversarial networks (GANs)

Pakpoom Singkorapoom : An Analysis of Deep Learning Model's Robustness against GANS-based Poisoning Attack in Medical Imaging. Advisor: Asst. Prof. SURONAPEE PHOOMVUTHISARN, Ph.D.

Deep learning revolutionizes healthcare, particularly in medical image classification, with its analysis performance aided by public architectures and transfer learning for pre-trained models. However, these models are vulnerable to adversarial attacks as they rely on learned parameters, and their unexplainable nature can make it challenging to identify and fix the root cause of the model after an attack. Given the increasing use of pre-trained models in life-critical domains like healthcare, testing their robustness against attacks is essential. Evasion and poisoning attacks are two primary attack types, with poisoning attacks having a broader range of poison sample-generating methods, making testing model robustness under them more critical than under evasion attacks. Poisoning attacks do not require an attacker to have a complete understanding to corrupt the model, making them more likely to occur in the real world. This work evaluates the robustness of the famous pre-trained models trained as binary classifiers under poisonous label attack. The attacks use GANs to generate mislabeled fake images and feed poison samples to the model in a black box manner. The amount of performance degradation using classification metrics evaluates the model's robustness. We found that ConvNeXt architecture is the most robust against this type of attack, suggesting that transformer architecture can be used to build a more robust deep-learning model.

Field of Study:	Statistics	Student's Signature .....
Academic Year:	2022	Advisor's Signature .....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้จะไม่สามารถเกิดขึ้นได้หากปราศจากการชี้แนะของ ผศ. ดร. สุรณพีร์ ภูมิวุฒิสารและขอขอบคุณเจ้าของชุดข้อมูล ดร. Muhammad Chowdhury และคณะที่นำชุดข้อมูลภาพเอกซเรย์ปอดออกมาเผยแพร่เพื่อประโยชน์ทางการวิจัย

ภาคภูมิ สิงขรภูมิ

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในช่วงทศวรรษที่ผ่านมาเราได้เห็นนวัตกรรมมากมายที่เกิดมาจากการใช้เทคโนโลยี deep learning และมีหลายภาคส่วนที่ได้รับประโยชน์จากพลังของเทคโนโลยีนี้ งานการแพทย์และสาธารณสุขก็เป็นอีกภาคส่วนหนึ่งที่ได้รับประโยชน์โดยตรงจากการนำเทคโนโลยี deep learning เข้ามาใช้แต่ทว่าในงานที่ต้องคำนึงถึงชีวิตของผู้คนเป็นหลักนั้นการนำเทคโนโลยี deep learning มาใช้ต้องมีการพิจารณาถึงความปลอดภัย (security) ของตัวเทคโนโลยีและความคงทน (robustness) ต่อภัยการโจมตีภายนอกด้วยเป็นอย่างยิ่ง ในปี 2013 งานวิจัยของ (Szegedy et al., 2013) ได้แสดงให้เห็นว่า deep learning model สามารถทำให้ถูกหลอกได้ด้วยสิ่งที่งานวิจัยเรียกว่า adversarial example ซึ่งเป็น data ที่ถูกดัดแปลงอย่างแยบยลเมื่อใช้เป็นอินพุตของ model จะทำให้ model ทำนายข้อมูลนี้ผิดไปจากสิ่งที่ควรจะเป็นด้วยความมั่นใจที่สูงทำให้เกิดเป็นความวิตกกังวลถึงการนำเทคโนโลยี deep learning การโจมตีแบบนี้ถูกเรียกว่า adversarial attack ในส่วนของงานทางการแพทย์และสาธารณสุขนั้นหนึ่งในรูปแบบของข้อมูลที่ถูกนำมาใช้สำหรับการสอน deep learning model มากที่สุดนั้นคือ ภาพทางการแพทย์ (Zhou et al., 2021) ซึ่งได้มาจากหลากหลายแหล่งเช่น จากการทำ x-ray, CT-scan, และ MRI เป็นต้น ภาพทางการแพทย์เหล่านี้เป็นข้อมูลชั้นดีที่ให้ข้อมูลทางด้านสรีระร่างกายทั้งภายในและภายนอก เป็นตัวช่วยการวินิจฉัยและการพยากรณ์โรคอย่างมาก (Qayyum et al., 2020) แต่ก็มีงานศึกษาที่กล่าวถึงการปรับแต่งภาพทางการแพทย์และนำมาใช้สำหรับ adversarial attack ใน deep learning model ถูกที่พัฒนาขึ้นมาเพื่อเป็นตัวช่วยแพทย์ (Finlayson et al., 2018)

Adversarial attack ต่อความปลอดภัยนั้นสามารถแบ่งได้ออกเป็นสองประเภท 1. Evasion attack และ 2. Poisoning attack (Biggio & Roli, 2018) โดยที่ evasion attack นั้นจะเกิดขึ้นในขั้นตอนการใช้ model ทำนายข้อมูลใหม่ (inference time) และจะมีผลแค่กับผลทำนายของ model เท่านั้นไม่ได้มีผลร้ายถึงตัว model แต่ทว่า poisoning attack ซึ่งจะเป็นการโจมตี ณ เมื่อทำ model training นั้นจะทำให้ model ที่ได้เป็น model ที่ไม่มีประสิทธิภาพเท่าที่ควร (suboptimal) เนื่องจากผู้โจมตีได้ใส่ข้อมูลที่ถูกดัดแปลงเข้าไป ณ ตอนที่ model กำลังเรียนรู้เป็นผลให้การเรียนรู้นั้นไม่ถูกต้องอีกทั้งผู้โจมตียังสามารถใช้หลากหลายเทคนิคในการดัดแปลงข้อมูลเพื่อสร้าง poison ขึ้นมาเช่น การใช้เทคโนโลยีที่ซับซ้อนอย่าง generative adversarial networks (Kasichainula et al., 2021; Liu et al., 2020; Shi et al., 2018; Yang et al., 2017) ที่ได้รับ

ความนิยมในการนำมาใช้สร้างข้อมูลขึ้นมาใหม่เพื่อทำ data augmentation สำหรับภาพทางการแพทย์ (Bhagat & Bhaumik, 2019; Kim et al., 2022; Kora Venu & Ravula, 2020; Zhang et al., 2018) การโจมตีชนิดนี้ยังสามารถทำได้ถึงแม้ว่าผู้โจมตีจะมีความรู้เกี่ยวกับเป้าหมายน้อยและสามารถใช้ประโยชน์จากความต้องการที่จะอัพเดท model เพื่อใส่ poison เข้าไปในระบบโดยที่ไม่จำเป็นต้องไปเข้าควบคุมระบบเป้าหมาย (Qayyum et al., 2020) และด้วยเหตุผลทั้งมวลเหล่านี้ทางผู้วิจัยมีความเห็นว่า poisoning attack นั้นเป็นสิ่งที่เราต้องให้ความคำนึงเป็นอย่างยิ่งมื่อต้องการจะพัฒนา deep learning model สำหรับงานด้านการแพทย์

ในวิทยานิพนธ์ฉบับนี้นทางผู้วิจัยได้ทำการศึกษาความคงทนของ deep learning model ที่ล้ำสมัยในงานการคัดแยกภาพทางการแพทย์ภายใต้การโจมตี poisoning attack เริ่มต้นทางผู้วิจัยได้ใช้กระบวนการโจมตีที่นำเสนอใน (Liu et al., 2021) ซึ่งเป็นการโจมตีแบบ black box โดยใช้ generative adversarial networks หรือ GANs เป็นหลักในการสร้าง poison และทำการสลับป้ายกำกับของข้อมูลที่ถูกสร้างขึ้นให้ไม่ถูกต้อง การโจมตีแบบ black box นั้นมีความสมจริงเนื่องจากผู้โจมตีสามารถมีความรู้เกี่ยวกับระบบเป้าหมายแค่น้อยนิดก็สามารถทำใหเกิดการโจมตีได้ หลังจากนั้นเราได้ทำการทำการทดลองจำลองการโจมตีกับ deep learning model ที่ล้ำสมัย 5 models ที่ถูกสร้างเพื่องานการจำแนกแบบไบนารีเพื่อหาสถาปัตยกรรมของ model ที่คงทนต่อการโจมตีมากที่สุดและน้อยที่สุด วิทยานิพนธ์ฉบับได้นำเสนอว่าแม้กระทั่ง model ที่มีสถาปัตยกรรมที่ล้ำสมัยนั้นก็ยังสามารถถูกโจมตีได้โดยเฉพาะในงานภาพทางการแพทย์และการค้นพบของงานวิจัยนี้จะเป็นตัวต่อยอดให้เกิดการสร้างสถาปัตยกรรมที่มีความคงทนต่อการโจมตีมากขึ้นเพื่อการใช้งานเทคโนโลยี deep learning ที่ปลอดภัยต่อชีวิตของผู้คน

1.2 วัตถุประสงค์

เพื่อศึกษาความคงทนของ deep learning model ที่มีสถาปัตยกรรมที่แตกต่างกันต่อการโจมตี poisoning attack และศึกษาความสามารถในการสร้าง poison ด้วยวิธีที่มีพื้นฐานมาจาก generative adversarial networks

1.3 ขอบเขตของการศึกษา

1. ทำการทดลองจำลองการโจมตี deep learning model ที่มีสถาปัตยกรรมแตกต่างกัน 5 แบบโดยใช้ชุดข้อมูลภาพทางการแพทย์เป็นภาพเอกซเรย์ปอดเป็นตัวแทนของภาพทางการแพทย์
2. ทำการทดลองจำลองการโจมตี deep learning model ที่มีสถาปัตยกรรมต่างกัน 5 แบบที่ถูกพัฒนาขึ้นมาสำหรับงานการจำแนกแบบไบนารีซึ่งเป็น model ประเภท supervised learning สำหรับตรวจหาโรคจากชุดข้อมูลภาพเอกซเรย์ปอด

3. ประเภทของ poisoning attack ที่ศึกษานั้นคือ black box และ accuracy drop attack ซึ่งการเป็นการโจมตีประเภทลดประสิทธิภาพของ model โดยที่ผู้โจมตีสามารถมีความรู้เกี่ยวกับระบบเป้าหมายน้อยนิดก็สามารถทำให้เกิดการโจมตีได้และวิธีการสร้าง poison จะใช้วิธี poisonous label attack ที่ถูกเสนอใน (Liu et al., 2021)

1.4 วิธีดำเนินการศึกษา

1. รวบรวมงานวิจัยที่เกี่ยวข้องกับ poisoning attack ที่ศึกษาใน deep learning model และในงานภาพทางการแพทย์
2. หาชุดข้อมูลที่เหมาะสมสำหรับการพัฒนา model สำหรับใช้ในการจำแนกภาพปอดแบบไบนารี
3. พัฒนา deep learning model ที่มีสถาปัตยกรรมแตกต่างกัน 5 แบบด้วยชุดข้อมูลที่ในกล่าวในข้อ 3
4. พัฒนา generative adversarial networks หรือ GANs ตาม (Liu et al., 2021) เพื่อใช้ในการสร้าง poison
5. ทำการทดลองจำลองการโจมตี deep learning model ที่ถูกพัฒนาขึ้นมาตามข้อ 3 จากนั้นบันทึกผลและสรุปผลการทดลอง

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. สามารถระบุหาและเปรียบเทียบประสิทธิภาพของ deep neural model ที่มีสถาปัตยกรรมแตกต่างกันที่มีความคงทนต่อการโจมตีได้เพื่อเป็นการนำร่องไปสู่การพัฒนาให้ได้ model ที่มีความปลอดภัยมากยิ่งขึ้นในงานที่มีต้องคำนึงถึงชีวิตอย่างงานทางการแพทย์และสาธารณสุข
2. รู้และเข้าใจถึงภัยอันตรายที่เกิดจากการนำ model ประเภท generative model ไปใช้ในทางที่ผิดเพื่อให้ตระหนักถึงการหาหนทางป้องกันการโจมตีที่อาจจะเกิดขึ้นจาก generative model และผลักดันให้มีการศึกษาวิจัยในเรื่องของความปลอดภัยของ deep learning model ในภาคของสุขภาพและการแพทย์ในประเทศไทย

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

เพื่อที่จะทำการทดลองจำลองการโจมตี deep learning model ด้วยวิธี poisoning attack ทางผู้วิจัยได้ศึกษาภาพรวมของ adversarial attack และตัวอย่างงานศึกษาที่มีการทดลองทำการโจมตี deep learning model ที่สร้างมาเพื่อจำแนกภาพทางการแพทย์และไม่ได้พัฒนามาเพื่องานทางการแพทย์ด้วยการโจมตีแบบ poisoning attack รวมถึงทฤษฎีที่เกี่ยวข้องกับ generative adversarial networks หรือ GANs ที่เป็นวิธีหลักในที่จะนำมาใช้สร้าง poison

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 Adversarial machine learning

Adversarial machine learning นั้นเป็นศาสตร์ที่ศึกษาทั้งในด้านความปลอดภัยความคงและความเป็นส่วนตัวของ machine learning model (Biggio & Roli, 2018) โดยในงานศึกษานั้นได้มีการบันทึกไว้ว่ามีการศึกษาครั้งแรกปี 2004 จากการค้นพบว่า email spam filter สามารถถูกหลอกด้วยวิธีที่คล้ายกับวิธีที่ถูกค้นพบกลับขึ้นมาอีกครั้งในปี 2013 ในงานศึกษาด้าน computer vision โดยหัวข้อหลักที่ศึกษาในศาสตร์ด้านนี้จะมีสามหัวข้อคือ 1. การสร้างการโจมตี 2. คิดค้นวิธีการประเมินความปลอดภัยและ 3. พัฒนาการป้องกัน โดยสาระสำคัญของการโจมตีนั้นคือการที่ผู้โจมตีจะสร้างหรือดัดแปลงเพื่อให้ได้ “ข้อมูลใหม่” ที่เมื่อใช้เป็นอินพุตของ model แล้วไม่ว่าจะเป็น ณ training หรือ ณ การนำ model ไปทำนายข้อมูลใหม่ทำให้ model นั้นประพฤติไม่ปกติหรือไม่เป็นไปตามที่ผู้พัฒนาออกแบบไว้

ตามหลักของการโมเดลภัยคุกคามแบบ proactive security ซึ่งเป็นการออกแบบความปลอดภัยที่เน้นลดระดับอันตรายที่จะเกิดขึ้นในอนาคตจากการโจมตีที่ไม่คาดคิด การออกแบบระบบ machine learning ให้ปลอดภัยต่อผู้โจมตีในควรคำนึงถึงปัจจัยต่อไปนี้

1. เป้าหมายของผู้โจมตี (Attacker’s goal) โดยสามรถนำมาวิเคราะห์ได้ 3 อย่าง

1.1 การละเมิดการละเมิดความปลอดภัยที่ผู้โจมตีต้องการ (desired security violation) โดยผู้โจมตีต้องการที่จะหลบหลีกการตรวจจับโดยที่ไม่ทำอะไรกับระบบ (integrity violation) หรือผู้โจมตีต้องการที่จะลดประสิทธิภาพของระบบเพื่อให้เกิดความผิดพลาดต่อผู้ใช้ (availability violation) หรือผู้โจมตีต้องการที่จะลอบล้วงข้อมูลลับเกี่ยวกับระบบ, ผู้ใช้ระบบ หรือข้อมูลในระบบ (privacy violation)

1.2 ความเจาะจงของการโจมตี (attack specificity) โดยผู้โจมตีสนใจที่จะทำให้เกิดการทำจำแนกผิด (misclassification) ในรูปแบบ “เฉพาะเจาะจง” (targeted) หรือ “ไม่เฉพาะเจาะจง” (indiscriminate) ต่อชุดของข้อมูลหนึ่งๆ

1.3 ความเจาะจงของความคลาดเคลื่อน (error specificity) โดยผู้โจมตีต้องการการทำนายผิดต่อข้อมูลหนึ่งๆไปเป็น class ที่เฉพาะเจาะจง (specific) หรือให้เกิดการทำนายผิดต่อข้อมูลหนึ่งไปเป็น class อะไรก็ได้ที่ไม่ใช่ ground truth (generic)

2. ความรู้ของผู้โจมตี (Attacker’s knowledge) ผู้โจมตีสามารถมีความรู้เกี่ยวกับระบบเป้าหมายได้หลายระดับโดยความรู้ในที่นี้อาจจะเป็นการที่ผู้โจมตีรู้ชุดข้อมูลหรืออัลกอริทึมที่ใช้หรือ loss function หรือแม้กระทั่งพารามิเตอร์ที่ได้หรือไฮเปอร์พารามิเตอร์ของระบบ โดยจะแบ่งระดับความรู้ของผู้โจมตีเป็นดังนี้

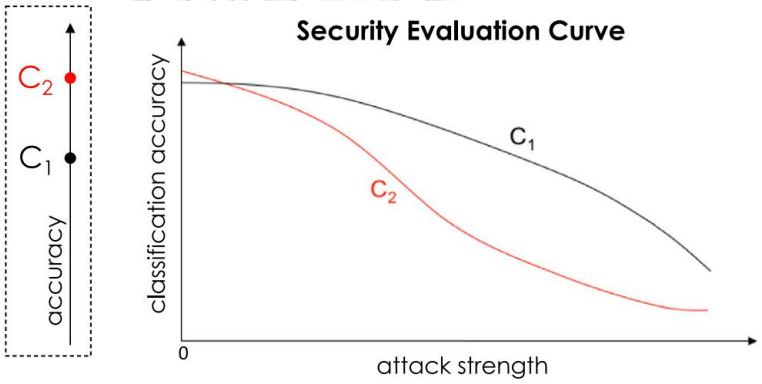
2.1 ความรู้แบบเสร็จสรรพ (Perfect knowledge, PK) ในกรณีนี้ผู้โจมตีจะรู้ข้อมูลทุกอย่างเกี่ยวกับระบบที่ตนต้องการจะโจมตีซึ่งเกิดขึ้นได้ยากในความเป็นจริงแต่การจำลองการโจมตีในรูปแบบนี้จะช่วยให้ผู้ออกแบบระบบสามารถวิเคราะห์ประสิทธิภาพได้ในสถานการณ์ที่เลวร้ายที่สุด (worst-case evaluation) การโจมตีในกรณีที่ผู้โจมตีมีความรู้เสร็จสรรพจะเรียกว่า white box attack

2.2 ความรู้แบบจำกัด (Limited Knowledge, LK) เป็นกรณีที่ผู้โจมตีมีความรู้เกี่ยวกับ feature และประเภทอัลกอริทึมที่ใช้แต่ไม่มีความรู้เกี่ยวกับชุดข้อมูลและพารามิเตอร์ที่ถูกประมาณหรือไฮเปอร์พารามิเตอร์ที่ใช้ ผู้โจมตียังสามารถใช้ความรู้เกี่ยวกับ feature ที่ machine learning model ใช้ในการไปเก็บรวบรวมชุดข้อมูลที่คล้ายคลึงกันมาใช้กับ model ตัวแทนเพื่อดูผลลัพธ์ที่ได้จากทำนายซึ่งจะสามารถทำให้ผู้โจมตีประมาณพารามิเตอร์ของ model ที่ต้องการจะโจมตีได้ ในกรณีที่ผู้โจมตีมีความรู้แค่ feature ผู้โจมตีสามารถใช้วิธีสร้างการโจมตีต่อ model ใดๆตามสมมติฐานของผู้โจมตีแล้วใช้คุณสมบัติ “การส่งต่อ” (transferability) ของข้อมูลใหม่ร้ายที่สร้างขึ้นไปทดสอบที่ machine learning model ที่ต้องการจะโจมตี การโจมตีในกรณีที่ผู้โจมตีมีความรู้จำกัดจะเรียกว่า gray-box attack

2.3 ไร้ข้อมูล (Zero knowledge, ZK) เป็นกรณีที่ใกล้เคียงกับความเป็นจริงที่สุดนั่นคือผู้โจมตีมีความรู้เกี่ยวกับระบบเป้าหมายเล็กน้อยเท่านั้นและเพื่อความกระจ่างชัดถึงแม้ชื่อของหัวข้อนี้จะเป็น “ไร้ข้อมูล” ก็ตามแต่ผู้โจมตียังสามารถมีความรู้เกี่ยวกับระบบเป้าหมายได้อยู่ ยกตัวอย่างเช่น machine learning model ที่ถูกสอนด้วยข้อมูลประเภทภาพ ผู้โจมตีจะยังสามารถรู้ได้ว่า feature คือ pixels ของภาพและด้วยความที่ machine learning model หนึ่งๆนั้นถูกสร้างขึ้นมาเพื่องานที่เฉพาะเช่น model ที่ไว้สำหรับจำแนกภาพปอดเป็นโรคกับภาพปอดที่ปกติ ผู้โจมตีก็จะรู้ว่า model

นั้นถูกสอนมาด้วยด้วยภาพปอดดังนั้นผู้โจมตียังคงสามารถใช้ความสามารถการส่งต่อได้เหมือใน LK เพียงแค่มีความรู้เกี่ยวกับระบบน้อยกว่า การโจมตีในกรณีแบบ ZK จะเรียกว่า black-box attack หมายเหตุในงานศึกษาบางงานไม่ได้แยกความรู้ของผู้โจมตีแบบความรู้จำกัดออกจากไร้ข้อมูลแต่มองว่าเป็นประเภทเดียวกันเลยได้

นอกจากการประเมินประสิทธิภาพของ model ภายใต้การโจมตีที่ผู้โจมตีมีระดับความรู้ที่แตกต่างกันแล้วนั้นผู้ออกแบบระบบ machine learning ควรจะต้องประเมินระบบเมื่อระดับของการโจมตีนั้นเพิ่มมากขึ้นด้วยยกตัวอย่างเช่นการเพิ่มจำนวน poison เข้าไปในชุดข้อมูล training set เมื่อทำการทดลองจำลองการโจมตีแบบ poisoning attack โดยสามารถระบบได้ด้วย security evaluation curve ซึ่งมีความจำเป็นสำหรับการวิเคราะห์รูปแบบการโจมตีและการป้องกันที่แตกต่างกันตามบริบทของการทดลอง

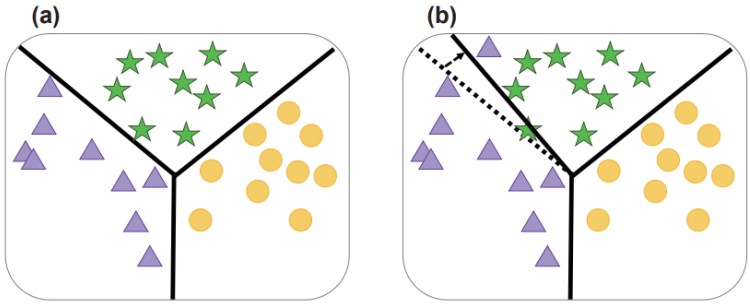


ภาพที่ 1 ภาพแสดง security evaluation curve (Biggio & Roli, 2018)

จากภาพที่ 1 จะเห็นได้ว่าความแม่นยำของตัวจำแนก c2 นั้นมากกว่าตัวจำแนก c1 แต่ภายใต้ adversarial attack ผู้ออกแบบระบบอาจจะต้องการ c1 มากกว่าเนื่องจากมีความคงทนมากกว่า c2 เมื่อระดับการโจมตีเพิ่มมากขึ้น

2.1.2 Poisoning attack

Poisoning attack เป็นประเภทของการจู่โจมที่ผู้โจมตีใส่ตัวอย่างข้อมูลที่ถูกดัดแปลงและถูกสร้างขึ้นมาอย่างพิถีพิถันเข้าไปชุดข้อมูล training set เพื่อรบกวนกระบวนการเรียนรู้ของ model เป็นผลให้ประสิทธิภาพของmodel ลดลงเมื่อนำ model ไปทำนายข้อมูลชุดใหม่โดยข้อมูลที่ถูกสร้างขึ้นมาในงานศึกษานี้จะเรียกว่า poison ซึ่งโดยทั่วไปจะมีคุณลักษณะคล้ายคลึงกับข้อมูลจริงเป็นอย่างมาก เช่น ถ้าในกรณีของภาพจะไม่สามารถแยกออกได้ว่าเป็นภาพที่ถูกสร้างขึ้นมาแต่ป่ายกำกับของข้อมูลจะผิดซึ่งเป็นความจงใจของผู้โจมตีทำให้เกิดการเปลี่ยนแปลงการกระจายตัวของข้อมูล (data distribution) เกิดขึ้น



ภาพที่ 2 ผลกระทบจาก poisoning attack ที่ทำให้เกิดการเปลี่ยนแปลงขอบเขตการตัดสินใจ (Liu et al., 2020)

งานศึกษาในปัจจุบันสามารถแบ่ง poisoning attack ออกเป็น 3 รูปแบบ (Liu et al., 2020) คือ

1. Accuracy drop attack ผู้โจมตีมีเป้าหมายที่จะลดความแม่นยำโดยรวมของ model ตอน inference time

2. Target misclassification attack ผู้โจมตีทำการเปลี่ยนแปลงข้อมูลในชุดข้อมูล test data โดยมีเป้าหมายคือบังคับให้ model ทำนายผลของตัวอย่างข้อมูลทดสอบที่เฉพาะเจาะจง (specific test instance) ผิดตอน inference time (Shafahi et al., 2018)

3. Backdoor attack ผู้โจมตีใส่สัญลักษณ์ (backdoor) บางอย่างเข้าไปในข้อมูลหรือ pre-trained model เพื่อควบคุม model ตอน inference time

Poisoning attack นั้นมีความเกี่ยวข้องกับงานทางการแพทย์มากกว่าการโจมตีตอน inference time อย่าง evasion attack เนื่องจากผู้โจมตีเพียงต้องการที่จะใส่ poison ที่สร้างขึ้นมาเข้าไปในชุดข้อมูลสำหรับสอน model ซึ่งไม่มีความจำเป็นที่จะต้องเข้าไปควบคุมข้อมูลโดยตรงอย่างที่

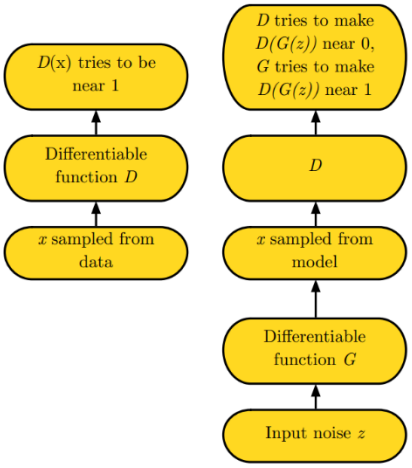
ทำใน evasion attack (Qayyum et al., 2020) อีกทั้งวิธีการสร้าง poison นั้นยังมีหลากหลายและมีผลกระทบกับ model ในระดับพารามิเตอร์ทำให้ model ผลลัพธ์นั้นไม่มีประสิทธิภาพเป็นไปตามที่ผู้พัฒนาต้องการในขณะที่การโจมตีอย่าง evasion attack นั้นส่งผลแค่กับผลทำนายของ model

2.1.3 Generative adversarial networks

Generative adversarial networks หรือ GANs (Ian J. Goodfellow et al., 2014) เป็นโครงสร้างหนึ่งของ deep learning model ที่จัดอยู่ในประเภทของ generative model ซึ่งเป็นประเภทของ model ที่พยายามจะประมาณการแจกแจงของข้อมูล (data distribution, p_{data}) โดยผลลัพธ์ที่ได้จากการประมาณจะแทนด้วย p_{model} (model distribution) GANs อยู่ในประเภทของ generative model ที่ไม่ได้ประมาณ probability distribution ของ p_{data} โดยตรงแต่จะใช้วิธีการสุ่มตัวอย่างการแจกแจงของโมเดลที่ถูกเรียนรู้มาเพื่อสังเคราะห์ข้อมูลขึ้นมาใหม่หรือเรียกว่า Implicit density model

GANs นั้นประกอบด้วย deep neural networks 2 ชุด คือ generator และ discriminator โดยกระบวนการ training ของ generator กับ discriminator นั้นสามารถอธิบายเป็นการเล่นเกมระหว่าง 2 ผู้เล่นโดย generator จะเป็นผู้สร้างข้อมูลปลอมขึ้นมาโดยจะสุ่ม noise ขึ้นมาจาก latent variable z ส่วน discriminator จะรับข้อมูลปลอมที่สร้างจาก generator และข้อมูลจริงจากชุดข้อมูล training set และจะเรียนรู้ที่จะแยกให้ออกว่าข้อมูลที่ได้รับเป็นข้อมูลจริงหรือข้อมูลที่สร้างขึ้นด้วย generator โดยเมื่อกระบวนการ training ผ่านไป generator จะต้องเรียนรู้ที่จะสร้างข้อมูลที่เหมือนกับถูกสุ่มขึ้นมาจากการแจกแจงของชุดข้อมูล training data

จากการที่เราสามารถอธิบายกระบวนการ training ของ GANs ได้ว่าเป็นการเล่นเกมระหว่าง 2 ผู้เล่นหรือ zero-sum game ซึ่งในอุดมคติแล้วเมื่อ generator สามารถที่จะสร้าง data ได้ราวกับว่าสุ่มมาจากการแจกแจงของชุดข้อมูล training data เมื่อนั้น discriminator จะทายผลว่าเป็นข้อมูลจริงหรือปลอมด้วยความน่าจะเป็นเท่ากับ 0.5 ณ จุดนั้นจะเรียกว่า GANs ลู่เข้าสู่ Nash equilibrium



ภาพที่ 3 ภาพรวมกระบวนการเฟรมเวิร์กของการพัฒนา generative adversarial networks (Goodfellow, 2016)

generator และ discriminator สามารถแทนได้ด้วยฟังก์ชันทางคณิตศาสตร์ที่สามารถหาอนุพันธ์ได้โดย generator จะแทนด้วยฟังก์ชัน G และ discriminator จะแทนด้วยฟังก์ชัน D กระบวนการ training ของ discriminator D จะเป็นไปตามรูปแบบทั่วไปของ supervised learning สำหรับงานจำแนกข้อมูลแบบไบนารีที่จะจำแนกอินพุตออกเป็น 2 class คือ จริงหรือปลอม โดย D จะรับ x จากชุดข้อมูล training data และเอาต์พุตจาก generator และมี θ^D เป็นพารามิเตอร์ของเน็ตเวิร์คส่วน generator G จะรับ noise z เป็นอินพุตและมี θ^G เป็นพารามิเตอร์ของระบบ ทั้ง generator และ discriminator จะมี cost function เป็นของตัวเอง โดย discriminator ต้องการจะลด $J^D(\theta^D, \theta^g)$ และจะทำเฉพาะเมื่อทำการสอน discriminator ซึ่งเหมือนกับ generator ที่ต้องการจะลด $J^G(\theta^D, \theta^G)$ และจะทำได้เฉพาะเมื่อทำการสอน generator และด้วยความที่ cost function ของแต่ละ network ขึ้นอยู่กับพารามิเตอร์ของฝ่ายตรงข้าม กระบวนการ training ของ GANs จะเหมาะสมที่จะอธิบายว่าเป็นการเล่นเกม (zero sum game) โดยที่ผลสรุปของเกมคือ Nash equilibrium ซึ่งในกรณีนี้คือ (θ^D, θ^G) ที่เป็น local minimum ของ J^D เทียบกับ θ^D และ local minimum ของ J^G เทียบกับ θ^G ใน parameter space สำหรับสมการ cost function ของ discriminator เป็นไปตามสมการด้านล่าง (Goodfellow, 2016)

$$J^D(\theta^D, \theta^G) = -\frac{1}{2}E_{x\sim p_{data}}[\log D(x)] - \frac{1}{2}E_z[\log(1 - D(G(z)]$$

ซึ่งคือ cross-entropy มาตรฐานที่ใช้สำหรับตัวจำแนกแบบไบนารีที่มี sigmoid activation function เป็น activation function ที่ output layer แต่แตกต่างกันตรงที่ discriminator จะถูกสอนด้วย mini-batch ของข้อมูล 2 ชุด คือข้อมูลจากชุดข้อมูล training data ที่จะมีป้ายกำกับเป็น 1 และข้อมูลที่มาจาก output ของ generator ($G(z)$) ที่จะมีป้ายกำกับเป็น 0 สำหรับการ training ของ GANs ในรูปแบบ minimax game หรือ zero-sum game นั้น cost function ของ generator เป็นไปตามสมการด้านล่าง

$$J^G = -J^D$$

ซึ่งสามารถสรุปได้เป็น value function ในรูปของ discriminator cost function ตามสมการด้านล่าง

$$V(\theta^D, \theta^G) = -J^D(\theta^D, \theta^G)$$

โดยที่ solution ของ minimax game เป็นไปตามสมการด้านล่าง

$$\theta^{G*} = \underset{\theta^G}{\operatorname{argmin}} \max_{\theta^D} V(\theta^D, \theta^G)$$

ตารางที่ 1 อัลกอริทึมการ training ของ generative adversarial networks

อัลกอริทึมที่ 1 Minibatch stochastic gradient descent training of generative adversarial nets. The number of steps to apply to the discriminator, ***k***, is a hyperparameter.

for number of training iterations ***do***

  for k steps ***do***

    Sample minibatch of m noise samples $\{z^{(1)}, \dots, z^{(m)}\}$ from noise prior $p_g(z)$

    Sample minibatch of m noise samples $\{x^{(1)}, \dots, x^{(m)}\}$ from data generating distribution $p_{data}(x)$

    Update discriminator by ascending its stochastic gradient

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} [logD\left(x^{(i)}\right) + \log\left(1 - D(G(z^i))\right)]$$

  end for

  Sample minibatch of m noise samples $\{z^{(1)}, \dots, z^{(m)}\}$ from noise prior $p_g(z)$

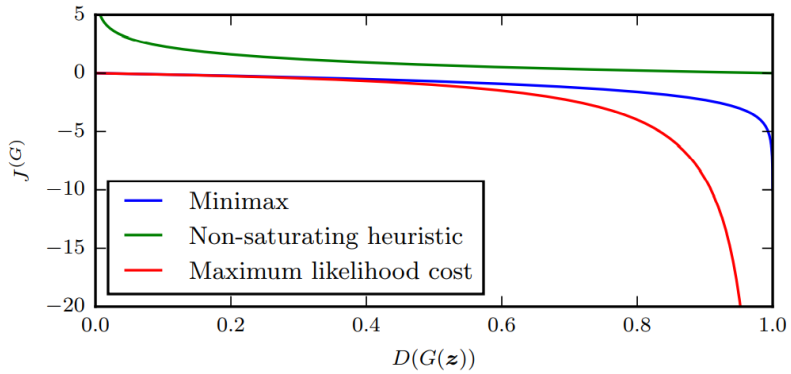
  Update generator by descending its stochastic gradient

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} [\log\left(1 - D(G(z^i))\right)]$$

end for

2.1.4 Non-saturating game (NS-GANs)

ในทางปฏิบัติการใช้ cost function ในรูปแบบของ minimax game นั้นจะทำให้เกิดปัญหา gradient vanishing ขึ้นกับ G เนื่องจากในช่วงแรกๆของการ training นั้น G ยังไม่สามารถเรียนรู้ที่จะสร้างเอาต์พุตให้เหมือนกับชุดข้อมูล training set ได้ทำให้ D สามารถแยกอินพุตที่ตนเองได้รับได้ ส่งผลให้ gradient ที่ส่งกลับไปหา G นั้นมีค่าน้อยมาก



ภาพที่ 4 ภาพแสดงปัญหา gradient vanishing ของ minimax GANs (Goodfellow, 2016)

แนวทางการแก้ไขคือการเปลี่ยน cost function ของ G เป็นดังนี้

$$J^G = -\frac{1}{2} E_z\left[logD\left(G(z)\right)\right]$$

ซึ่ง cost function ด้านบนนี้ยังคงเป็น cross-entropy เพียงแต่สลับป้ายกำกับของข้อมูลในที่นี้หมายถึงเอาต์พุตที่ได้จาก G จะให้มีป้ายกำกับเป็นข้อมูลจริง ในทางการตีความคือใน minimax game G ต้องการที่จะ minimize ความน่าจะเป็นที่ D จะใส่ป้ายกำกับให้กับเอาต์พุตของตัวเองถูก แต่ใน non saturating game นี้ G จะ maximize ความน่าจะเป็นที่ D ใส่ป้ายกำกับผิด (คือ ใส่ป้ายกำกับข้อมูลจริงให้เอาต์พุตที่มาจาก G) แต่เนื่องจากตอนนี้การใช้สมการในรูปแบบนี้ทำให้ไม่สามารถอธิบายเกมด้วย value function ตัวเดียวได้และเกมจะไม่ใช่ zero-sum game ทำให้ทฤษฎีที่พิสูจน์ถึงการลู่เข้าของ minimax game ไม่สามารถนำมาใช้กับ non-saturating game ได้

2.1.5 Reconstruction loss ใน GANs

โดยปกติแล้ว Reconstruction loss จะถูกใช้ใน auto encoder แต่ก็ได้มีงานศึกษาที่นำ L_1 reconstruction loss มาเพิ่มข้าไปใน cost function ของ G และ D และสามารถช่วยแก้ปัญหาการ training ที่ไม่มั่นคงและ mode collapse ของ GANs ได้รวมถึงเพิ่มคุณภาพของข้อมูลที่ถูกสร้างขึ้นด้วย G (Isola et al., 2017; Li et al., 2019) โดยสมการของ L_1 reconstruction loss อย่างง่ายเป็นดังนี้

$$L_{L_1}(G) = \lambda E_{x,y\sim p_{data},\ z\sim p_z(z)}[\|x - G(z)\|_1]$$

โดยที่

λ คือค่าน้ำหนักหรือค่าสัมประสิทธิ์ reconstruction

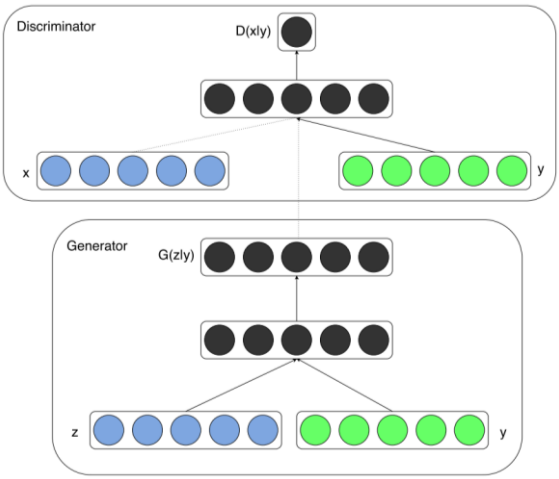
x คือข้อมูลจริงจากชุดข้อมูล training set

$G(z)$ คือเอาต์พุตจาก generator

2.1.6 Conditional generative adversarial networks (cGANS)

ใน GANs ดั้งเดิมถึงแม้ว่าจะสามารถทำการ training ระบบให้ลู่เข้าได้แล้วแต่ว่าเอาต์พุตที่สร้างขึ้นมากจาก G นั้นเป็นแบบสุ่มไม่สามารถควบคุมได้ตามที่ต้องการได้ ในปีเดียวกับที่ GANs ดั้งเดิมถูกนำเสนอ (Mirza & Osindero, 2014) ได้เสนอรูปแบบของ GANs ที่สามารถควบคุมการสร้างข้อมูลใหม่ของ G ได้ด้วยการใส่ข้อมูลเพิ่มเติม (extra information, y) เข้าไปกับทั้งอินพุตของ G และ D โดยข้อมูลเพิ่มเติม y นั้นสามารถเป็นอะไรก็ได้ในทางทฤษฎีเช่นป้ายกำกับหรือข้อมูลจากโดเมนอื่น objective function ของ minimax gam เป็นไปตามสมการด้านล่าง

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}}[\log D(x|y)] + E_{z\sim p_z(z)}\left[\log\left(1 - D\left(G(z|y)\right)\right)\right]$$



ภาพที่ 5 แผนผังแสดงโครงสร้าง conditional GANs อย่างง่าย (Mirza & Osindero, 2014)

2.1.7 Deep convolutional generative adversarial networks (DCGAN)

งานวิจัยของ (Radford et al., 2015) เป็นงานแรกที่ประสบความสำเร็จในความพยายามที่จะเพิ่มประสิทธิภาพของ GANs ด้วยการใช้ convolution neural networks งานวิจัยชิ้นนี้ได้มาจากการทำการทดลองค้นคว้าอย่างหนักจนเกิดเป็นต้นแบบวิธีของการโอน GANs โดยที่ G และ D มีโครงสร้างที่เป็น convolutional neural networks โดยผู้จัดทำตั้งชื่อโครงสร้างนี้ใหม่ว่า DCGAN และต่อมาโครงสร้างนี้ได้กลายเป็นมาตราฐานในการสร้าง GANs จนถึงปัจจุบัน สิ่งที่คณะผู้จัดทำค้นพบมีดังนี้ สำหรับกระบวนการ training DCGAN ที่มีเสถียรภาพมีดังนี้

คำแนะนำสำหรับโครงสร้างของ G และ D

1. ไม่มีการใช้ deterministic pooling layers (เช่น max pooling operation) แต่เปลี่ยนเป็น strided convolution (convolution operation ที่มีการใช้ค่า stride มากกว่า 1) ใน D และ fractional-strided convolution หรือที่รู้จักกันในชื่อ deconvolution ใน G

2. ใช้ batch normalization ทั้งใน G และ D

3. ไม่มีการใช้ fully connected architecture

4. ใช้ ReLU activation function ใน G ยกเว้นสำหรับ output ให้ใช้ Tanh

5. สำหรับ activation function ใน D ให้ใช้ LeakyReLU

คำแนะนำ hyper-parameters ที่ใช้สำหรับการ training คณะผู้จัดทำได้ทำการทดลองโครงสร้างตนเองเสนอขึ้นมากับ dataset ทั้งหมด 3 dataset ประกอบไปด้วย LSUN bedrooms, Imagenet-1k, Faces โดยองค์ประกอบของการ training มีดังนี้

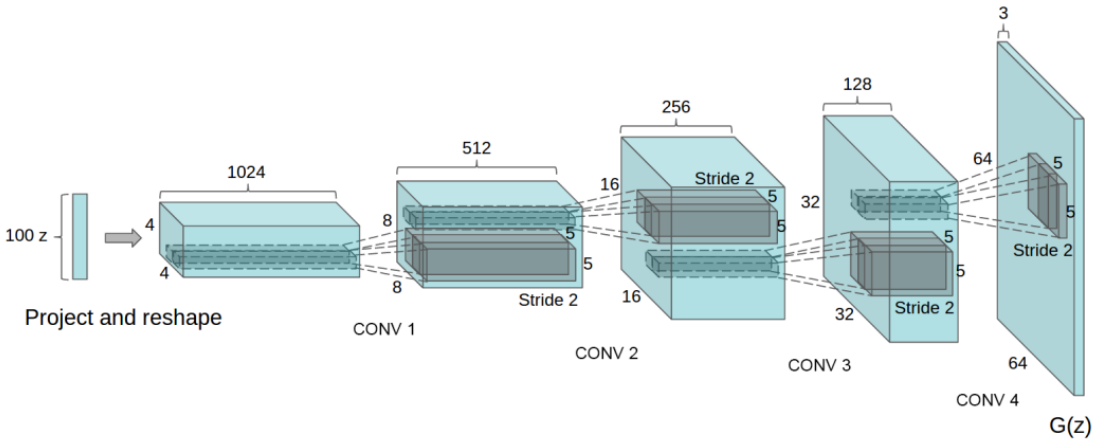
1. เปลี่ยนค่า pixel ของ input จาก training data ให้อยู่ในช่วง $[-1, 1]$

2.ใช้ batch size ขนาด 128

3.weights ใน neural networks เริ่มต้นให้สุ่มมาจาก zero-centered normal distribution ที่มีค่าเบี่ยงเบนมาตรฐาน 0.02

4.ใน LeakyReLU activation function ให้ใช้ค่า slope เท่ากับ 0.2

5.ใช้ Adam optimizer กับ learning rate ค่า 0.0002 และเปลี่ยนค่า momentum β_1 ให้เป็น 0.5



ภาพที่ 6 โครงสร้างแบบ all convolution style (Radford et al., 2015)

2.1.8 Fréchet Inception Distance (FID)

FID เป็นปริมาณที่ถูกเสนอโดย (Heusel et al., 2017) เป็นปริมาณที่ถูกปรับปรุงมาจาก inception score (IS) โดยการใช้ค่าสถิติของภาพจริงและภาพที่ถูกสร้างจาก generator มาใช้หาค่าความเหมือน และถูกใช้อย่างแพรหลายในการวัดคุณภาพของข้อมูลที่ถูกสร้างขึ้นมาโดย generative model โดยค่า FID ที่ต่ำจะยิ่งบ่งบอกถึงคุณภาพของข้อมูลจาก generator ที่ใกล้เคียงกับข้อมูลจริง การหาค่า FID เป็นไปตามสมการด้านล่าง

$$d^2(x,g) = \left\|\mu_x - \mu_g\right\|^2 + Tr\left(\Sigma_x + \Sigma_g - 2\left(\Sigma_x\Sigma_g\right)^{\frac{1}{2}}\right)$$

โดยที่

μ_x, μ_g คือเวกเตอร์ค่าเฉลี่ยของข้อมูลจริงและข้อมูลสังเคราะห์ที่สร้างจาก generator

Σ_x, Σ_g คือเมทริกซ์ความแปรปรวนร่วม (covariance matrix) ของข้อมูลจริงและข้อมูลสังเคราะห์ที่สร้างจากgenerator

$Tr()$ คือ trace ของเมทริกซ์

และ $X_x \sim N(\mu_x, \Sigma_x), X_g \sim N(\mu_g, \Sigma_g)$ คือเวกเตอร์ที่ออกมาจากเลเยอร์สุดท้ายของ inception v3 ซึ่งจะมีขนาด 2048 ของข้อมูลจริงและข้อมูลสังเคราะห์ที่สร้างจาก generator

2.2 งานวิจัยที่เกี่ยวข้อง

อย่างที่ได้กล่าวไปแล้วในหัวข้อที่ผ่านมา งานศึกษาในปัจจุบันนั้นได้แบ่ง adversarial attack ออกเป็นสองแบบแยกตามจุดเวลาที่โจมตีเป็น evasion attack และ poisoning attack ใน evasion attack นั้นปรับเปลี่ยนข้อมูลอินพุตสู่ model ตอน inference time เป็นผลนำไปสู่การทำนายที่ผิดพลาดยกตัวอย่างเช่น งานศึกษาของ (Ian J Goodfellow et al., 2014) ได้แสดงให้เห็นว่า GoogLeNet นั้นได้ทำนายรูปหมีแพนด้าที่ถูกก่อกวนออกมาเป็นชะนีด้วยความมั่นใจที่สูง สำหรับ poisoning attack ที่โจมตี ณ training time หรือ update time ผู้โจมตีจะสร้างข้อมูลใหม่ขึ้นมาพร้อมกับป้ายกำกับที่ไม่ถูกต้องเมื่อนำไปสอน model หรืออัพเดท model นั้นจะทำให้เกิดการเปลี่ยนแปลงขอบเขตการตัดสินใจ

ในการใช้เทคโนโลยี deep learning ในงานทางการแพทย์และสาธารณสุขนั้น poisoning attack มีความเกี่ยวข้องที่ต้องระวังเป็นอย่างมากเนื่องจากการโจมตีด้วยวิธีนี้นั้นอาศัยแค่ต้องการให้ผู้โจมตีปล่อยข้อมูล poison เข้าสู่ระบบด้วยวิธีใดก็ได้ซึ่งถ้าเทียบกับ evasion attack ที่ต้องมีการเข้าถึงข้อมูลเพื่อใส่ความเปลี่ยนเข้าไปนั้น poisoning attack จะสามารถทำให้เกิดขึ้นได้จริงมากกว่า นอกจากนี้ poisoning attack นั้นมีผลต่อ model ในระดับพารามิเตอร์เนื่องจากทำให้ขอบเขตการตัดสินใจนั้นเปลี่ยนไปไม่ใช่แค่มีผลต่อผลทำนายแบบการโจมตีตอน inference time แบบ evasion attack

ปัจจุบันได้มีงานศึกษาจำนวนหนึ่งที่ทำการศึกษา poisoning attack ใน deep learning model งานวิจัยของ (Muñoz-González et al., 2017) ผู้เขียนได้ทำการทดลองกับ deep learning model ที่พัฒนามาเพื่องานตรวจจับมัลแวร์และงานจำแนกภาพลายมือตัวเลข โดยผู้เขียนได้เสนอวิธีการสร้าง poison ที่มีชื่อว่าวิธี back gradient ที่สามารถเกาะรอยคำนวณย้อนกลับการหาอนุพันธ์ของ loss function เพื่อนำมาใช้แก้สมการ optimization สองระดับและได้ผลลัพธ์เป็น poison ในงานตรวจจับมัลแวร์ผู้เขียนได้แสดงให้เห็นว่าอัตราการผิดพลาดที่ชุดข้อมูล test set ของ neural network 1 เลเยอร์นั้นที่มี 10 neurons นั้นเพิ่มขึ้นจากปกติถึง 25% ในการทดลองแบบ white box แต่ทว่าสำหรับ model จำแนกภาพลายมือตัวเลขนั้นอัตราการผิดพลาดที่ชุดข้อมูล test set นั้นเพิ่มขึ้นแค่เล็กน้อยโดยมี model เป้าหมายเป็น convolutional neural networks นอกจากนั้นวิธีการ back gradient นั้นใช้ทรัพยากรณ์และเวลามหาศาลเพื่อที่จะสร้าง poison จึงไม่สมจริง งานศึกษาใน (Yang et al., 2017) ได้นำวิธี back gradient มาปรับปรุงและสามารถลดเวลาการสร้าง poison ไปได้ 200 เท่าโดยการนำสถาปัตยกรรมของ generative adversarial networks มาใช้โดยมี autoencoder เป็น generator และ model เป้าหมายเป็น discriminator ผู้เขียนได้ทำการทดลองกับชุดข้อมูล cifar-10 และ mnist ด้วยมี neural networks 2 เลเยอร์และ LeNet เป็น model เป้าหมายโดยมีผลการทดลองคือสามารถเพิ่มอัตราการผิดพลาดได้อย่างน้อย 16.59 % กับ

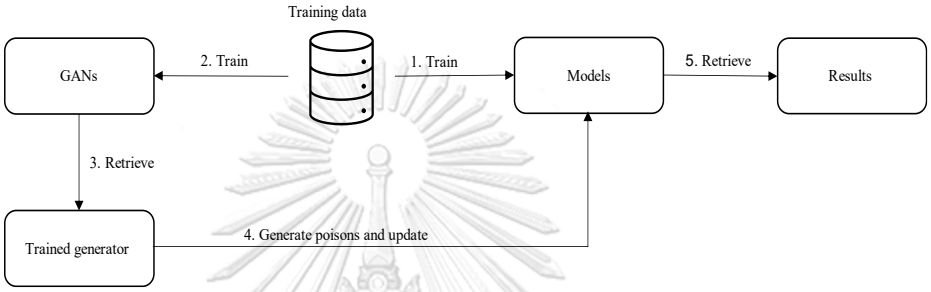
neural networks 2 เลเยอร์ที่สอนด้วยชุดข้อมูล mnist และ 20.74 % กับ LeNet ที่สอนด้วยชุดข้อมูล cifar-10 ถึงแม้ว่างานวิจัยที่กล่าวไปข้างต้นจะสามารถโจมตี model ได้แต่ว่าผู้เขียนได้ทำการทดลองในรูปแบบ white box attack ซึ่งไม่สมจริง งานวิจัย (Liu et al., 2021) ผู้เขียนได้เสนอวิธีการโจมตีชื่อว่า poisonous label attack ซึ่งใช้ DCGAN กับ reconstruction loss ในการสร้าง poison ที่ดูแล้วสมจริงและสร้างป้ายกำกับที่ผิดด้วยวิธี probability transition vector ผู้เขียนได้ทำการทดลองโจมตี LeNet ที่ถูกสอนด้วยชุดข้อมูล mnist และสามรถลดค่าความแม่นยำของ model ที่ชุดข้อมูล test set ได้ถึง 65.4 % เมื่อใส่ poison เข้าไปในระบบ 900 ตัว แต่เมื่อผู้เขียนได้เปรียบเทียบวิธีตัวเองกับสองวิธีก่อนหน้าที่จำนวน poison เท่ากันนั้นจะได้ผลว่าวิธีก่อนหน้าสองวิธีสามารถลดควาแม่นยำของ model ที่ชุดข้อมูล test set ได้มากกว่าแต่ทว่าวิธี poisonous label attack นั้นมีข้อดีที่สองวิธีก่อนหน้าไม่มีนั่นคือการที่วิธีนี้เป็นการโจมตีแบบ black box ซึ่งมีความสนใจและสามารถทำให้เกิดขึ้นได้ง่ายกว่า

จะเห็นได้ว่ามีงานวิจัยจำนวนหนึ่งได้ศึกษาและเสนอวิธีการโจมตี deep learning model และสามารถโจมตีได้แต่ทว่างานทดลองของงานเหล่านั้นทำการทดลองแต่กับชุดข้อมูลที่เป็นมาตราฐาน ไม่ใช่ชุดข้อมูลที่จะนำมาใช้ในโลกความจริงและก็ไม่ใช่ชุดข้อมูลสำหรับงานด้านการแพทย์อีกด้วย นอกจากนี้สถาปัตยกรรมของ model เป้าหมายยังเป็นแบบพื้นฐาน (Mozaffari-Kermani et al., 2015) ได้เสนอการโจมตี poisoning แบบที่ไม่เลือก model โดยการใช้สถิติของชุดข้อมูล training set หรือชุดข้อมูลตัวแทน โดยทำการทดลองกับชุดข้อมูลทางการแพทย์แบบตาราง 5 ชุดและกับ model เป้าหามาย 6 models ซึ่งรวมถึง deep learning model โดยได้ผลลัพธ์คือเมื่อใส่ poison เข้าไป 30 % เทียบกับจำนวนชุดข้อมูล training set ทั้งหมดสามารถลดความแม่นยำที่ชุดข้อมูล test set ของ model ลงมาได้ 20 % และ 26 % สำหรับ deep learning model ที่สอนด้วยชุดข้อมูลโรคไทรอยด์และโรคมะเร็งเต้านม งานวิจัยของ (Finlayson et al., 2018) ได้อธิบายว่าทำไม deep learning model ในงานด้านการแพทย์และสาธารณสุขนั้นถึงมีความเสี่ยงที่จะถูกโจมตีในแง่ของแรงจูงใจทางการเงินและปัญหาทางเทคนิคของระบบในโรงพยาบาล ผู้เขียนได้กล่าวถึงปัญหา ground truth ในการวินิจฉัยโรคในงานการจำแนกภาพทางการแพทย์นั้นครุมเครือต้องอาศัยนักรังสีวิทยาที่มีประสบการณ์เท่านั้นและคนเหล่านี้หาตัวได้ยากหรือค่าตัวแพงเพราะฉะนั้นถ้าเกิดการปรับเปลี่ยนข้อมูลภาพทางการแพทย์เกิดขึ้นอย่างจงใจจะสามารถตรวจจับได้ยากอีกทั้งภาพทางการแพทย์นั้นยังมีความเป็นมาตราฐานที่สูงไม่ได้มีความหลากหลายเหมือนภาพทั่วไปเนื่องจากภาพเหล่านี้ถูกถ่ายด้วยระบบที่ถูกตั้งค่ามาอย่างดีไม่ว่าจะเป็นเรื่องตำแหน่งหรือค่าความเข้มของแสงซึ่งความเป็นมาตรฐานสูงนี้เป็นอีกหนึ่งปัจจัยที่จะทำให้เกิดการโจมตีได้ง่ายขึ้นเนื่องจาก poison ที่สร้างขึ้นมาจะถูกตรวจจับได้ยากและความหลากลหลายของภาพทั่วไปนั้นสามารถลดผลของการโจมตีได้ ผู้เขียนได้ทำการทดลองโดยมี model เป้าหมายคือ ResNet50 และใช้ข้อมูลสามชุดนั่นคือ ชุดข้อมูลภาพเบาหวานขึ้นตา ชุด

ข้อมูลภาพเอกซเรย์ปอดบวม และภาพถ่ายมะเร็งผิวหนัง และรายงานว่า ResNet50 นั้นสามารถถูกหลอกได้โดยสิ้นเชิง (Asgari Taghanaki et al., 2018) ได้ทำการทดลองแบบเดียวกับ (Finlayson et al., 2018) โดยใช้การโจมตีทั้งในรูปแบบ white box และ black box โดยผู้เขียนได้ทำการทดลองกับ model เป้าหมายคือ Inception-ResNet-v2 และ Nasnet-Large และได้ผลลัพธ์เหมือนกันงานวิจัยของ (Asgari Taghanaki et al., 2018; Finlayson et al., 2018) นั้นทำให้เห็นว่า deep learning model ที่มีสถาปัตยกรรมที่สมัยใหญ่ที่พัฒนามาเพื่อจำแนกภาพทางการแพทย์สามารถถูกลดทอดประสิทธิภาพได้แต่ทว่าทั้งสองงานวิจัยนั้นเลือกใช้ evasion attack ซึ่งทางผู้วิจัยเชื่อว่ามีประสิทธิภาพด้อยกว่า poisoning attack ในงานทางด้านการแพทย์และสาธารณสุข วิทยานิพนธ์ฉบับนี้จึงศึกษาผลการโจมตีของ poisoning attack ที่ออกแบบมาเพื่อลดประสิทธิภาพโดยรวมด้วยวิธีการโจมตีแบบ black box หรือ poisonous label attack ผู้วิจัยทำการทดลองโดยใช้ deep learning model สำหรับการจำแนกแบบไบนารีที่มีสถาปัตยกรรมสมัยใหม่ 5 models กับชุดข้อมูลภาพเอกซเรย์

บทที่ 3
วิธีการดำเนินการวิจัย

บทนี้จะกล่าวถึงขั้นตอนการการทดลองจำลองการโจมตีด้วย poisoning attack โดยอธิบายชุดข้อมูลที่ถูกนำมาใช้ การสอน deep learning model ที่ใช้เป็นเป้าหมายการโจมตี การสอน generative adversarial networks และการนำ generator ไปใช้ในการโจมตี poisonous label attack และอัลกอริทึมที่ใช้ในการโจมตี โดยภาพรวมของการทดลองเป็นดังนี้



ภาพที่ 7 ภาพรวมกระบวนการทำวิจัย

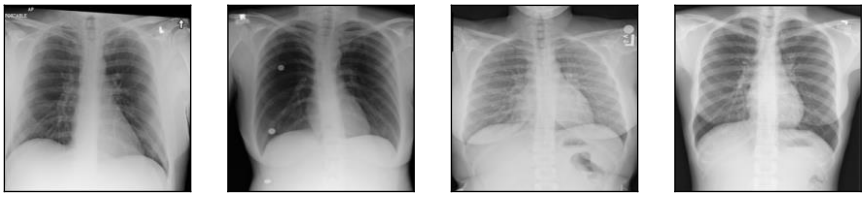
3.1 ชุดข้อมูล

ผู้วิจัยได้ใช้ชุดข้อมูลภาพเอกซเรย์ปอดแบบภาพระดับสีเทาจาก (Tahir et al., 2021) โดยชุดข้อมูลนี้ประกอบไปด้วยภาพสามชนิดคือ ภาพเอกซเรย์ปอดติดเชื้อโควิด19 ภาพเอกซเรย์ปอดบวม และภาพเอกซเรย์ปอดปกติ เนื่องจากขอบเขตของงานวิจัยคือการสอน deep learning model สำหรับการจำแนกแบบไบนารีเราจึงเลือกที่จะใช้ภาพเอกซเรย์ปอดบวมและภาพเอกซเรย์ปอดปกติเนื่องจากเป็นโรคปอดที่เกิดขึ้นมาก่อนนานแล้วโดยการกระจายตัวของภาพในแต่ละชุดข้อมูลเป็นไปตามตารางดังนี้

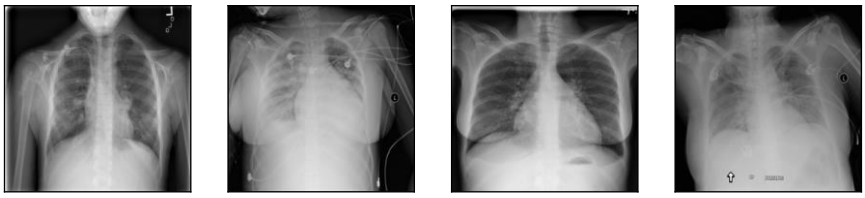
ตารางที่ 2 ลายละเอียดของชุดข้อมูล

ชื่อชุดข้อมูล	ประเภทของข้อมูล	จำนวนข้อมูล	ป้ายกำกับ
ชุดข้อมูล training set	ภาพเอกซเรย์ปอดปกติ	6849 ภาพ	0
	ภาพเอกซเรย์ปอดบวม	7208 ภาพ	1
ชุดข้อมูล validation set	ภาพเอกซเรย์ปอดปกติ	1712 ภาพ	0
	ภาพเอกซเรย์ปอดบวม	1802 ภาพ	1
ชุดข้อมูล test set	ภาพเอกซเรย์ปอดปกติ	2140 ภาพ	0
	ภาพเอกซเรย์ปอดบวม	2253 ภาพ	1

ภาพทุกภาพในแต่ละชุดข้อมูลได้ถูกทำความสะอาดและแปรรูปให้มีขนาด 128×128 พิกเซล และในแต่ละชุดข้อมูลจำนวนภาพมีความสมดุล



ภาพที่ 8 ตัวอย่างภาพเอกซเรย์ปอดปกติจากชุดข้อมูล training set



ภาพที่ 9 ตัวอย่างภาพเอกซเรย์ปอดบวมจากชุดข้อมูล training set

3.2 การสอน model เป้าหมาย

เพื่อที่จะประเมินความคงทนของ deep learning model ในลำดับแรกผู้วิจัยได้เลือก deep learning model ที่มีสถาปัตยกรรมล้ำสมัยทั้งหมด 5 รูปแบบมาพัฒนาสำหรับงานจำแนกภาพเอกซเรย์ปอดแบบไบนารีโดย deep learning model ที่เรียกมานั้นมีสถาปัตยกรรมดังต่อไปนี้

1.VGG16

2.ResNet50v2

3.MobileNetv2

4.Inceptionv3

5.ConvNext-Tiny

deep learning model เหล่านี้ล้วนมีความแตกต่างออกจากันทั้งในแง่ของจำนวนพารามิเตอร์ จำนวนเลเยอร์เทคนิคที่ใช้ในการออกแบบสถาปัตยกรรมเพราะฉะนั้นจึงควรจะมีมีพฤติกรรมที่แตกต่างกันเมื่ออยู่ภายใต้การโจมตี ผู้วิจัยทำการสอน model ทั้งหมดด้วยวิธีที่เหมือนกันเพื่อควบคุมการทดลองและใช้เทคนิค transfer learning ผู้วิจัยได้ใช้พารามิเตอร์ที่ถูกเรียนรู้มาแล้วกับชุดข้อมูล ImageNet เป็นพารามิเตอร์ตั้งต้นและปิดการอัพเดทพารามิเตอร์ส่วนนี้ไว้จากนั้นแทนที่ส่วนจำแนก

ตัวเก่าด้วยส่วนจำแนกใหม่โดยอ้างอิงมาจาก (Kora Venu & Ravula, 2020) โดยมีเลเยอร์ดังตารางด้านล่าง

ตารางที่ 3 ตารางแสดงองค์ประกอบของส่วนจำแนกสำหรับแต่ละสถาปัตยกรรม

ชื่อเลเยอร์	ค่าไฮเปอร์พารามิเตอร์ที่ใช้
Dense	512 nodes
Batch normalization	
ReLU activation function	
Dropout	อัตรา drop out 0.2
Dense	128 nodes
Batch normalization	
ReLU activation function	
Dropout	อัตรา drop out 0.2
Dense	64 nodes
Batch normalization	
ReLU activation function	
Dropout	อัตรา drop out 0.2
Dense	1 nodes

และทำการ training ด้วยไฮเปอร์พารามิเตอร์ดังนี้

ตารางที่ 4 ตารางแสดงค่าไฮเปอร์พารามิเตอร์สำหรับพัฒนา model เป้าหมาย

ไฮเปอร์พารามิเตอร์	ค่า
ตัวเลือก optimizer	Adam
ค่า learning rate	0.0001
ค่า epoch	300 รอบ
ค่า batch size	64
กลยุทธ์ early stopping	ใช้ที่ 16 รอบ

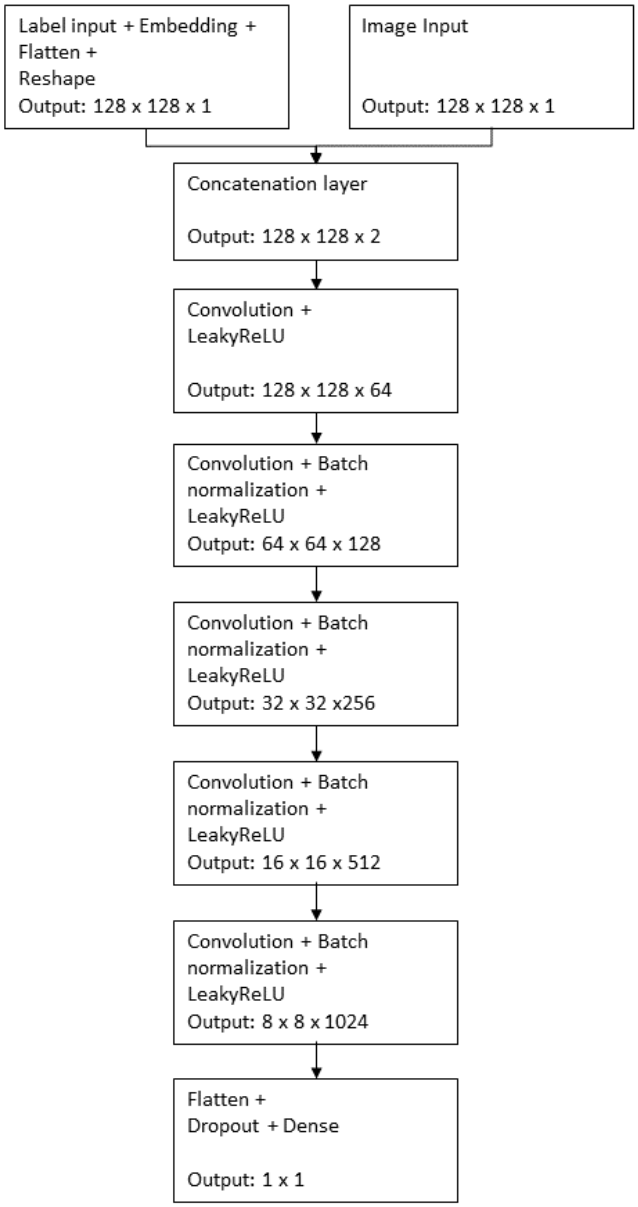
เมื่อทำการ training เสร็จเราได้ทำการ fine tuning ทุก model ต่อโดยใช้กลยุทธ์ตาม โดยเปิดการอัพเดทของพารามิเตอร์ทุกตัวในทุกเลเยอร์ของส่วนที่ถูกสอนมาด้วยชุดข้อมูล ImageNet แล้วทำการ training ต่ออีก 16 รอบด้วย learning rate ที่ต่ำกว่าเดิมและเลือกเฉพาะ model ที่ได้ประสิทธิภาพที่ชุดข้อมูล validation set สูงสุดในแต่ละสถาปัตยกรรมไปใช้ต่อโดยใช้ไฮเปอร์พารามิเตอร์ตามตารางดังนี้

ตารางที่ 5 ตารางแสดงค่าไฮเปอร์พารามิเตอร์สำหรับ fine tuning

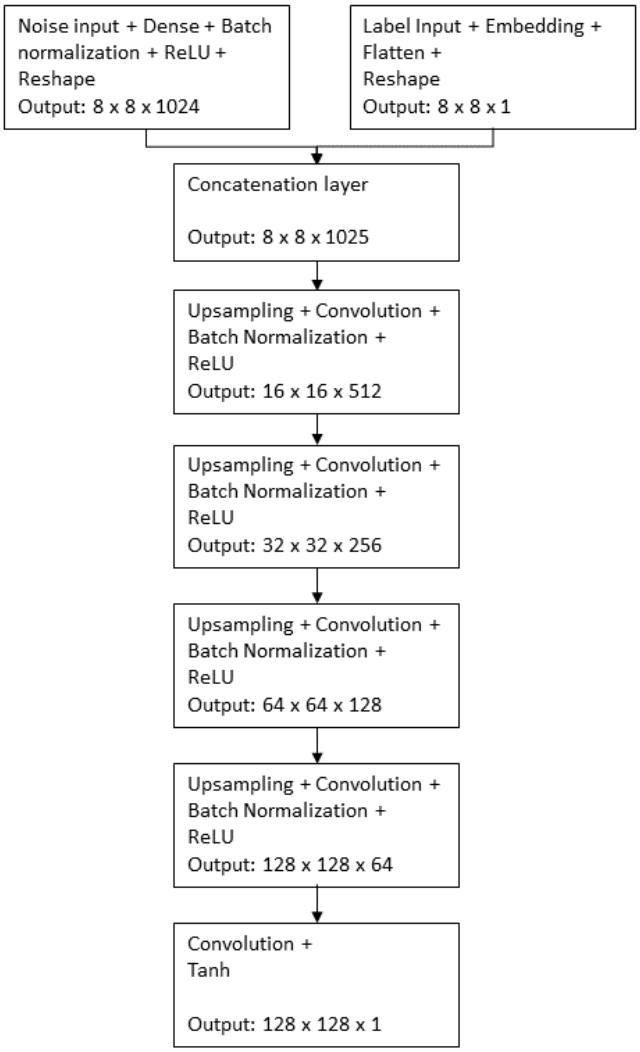
ไฮเปอร์พารามิเตอร์	ค่า
ค่า learning rate	0.00001
ค่า epoch	16 รอบ

3.3 การพัฒนา GANs

สร้างสถาปัตยกรรมของ generator และ discriminator ตามรูปแบบของ DCGAN แบบ conditional ที่ได้กล่าวไปใน 2.1.6 และ 2.1.7 โดยมีองค์ประกอบตามตารางด้านล่าง



ภาพที่ 10 แผนผังแสดงโครงสร้างของ discriminator



ภาพที่ 11 แผนผังแสดงโครงสร้างของ generator

ในส่วนของการ training นั้นผู้วิจัยได้ใช้เพียงชุดข้อมูล training set สำหรับการสอนเพราะว่าไม่ต้องการให้เกิดการรั่วไหลของข้อมูลสำหรับชุดข้อมูลที่ใช้ในการประเมิน model ทำการ training โดยใช้ไฮเปอร์พารามิเตอร์และเทคนิคเพิ่มเติมตามตารางด้านล่าง

ตารางที่ 6 ตารางแสดงค่าไฮเปอร์พารามิเตอร์สำหรับการพัฒนา GANs

ไฮเปอร์พารามิเตอร์	ค่า
ตัวเลือก loss function	Binary cross entropy + Reconstruction loss
ตัวเลือก optimizer	Adam
ค่า β_1 ของ Adam	0.5
ค่า β_2 ของ Adam	0.9
ค่า learning rate สำหรับ generator	0.0001
ค่า learning rate สำหรับ discriminator	0.0003
ค่า epochs	200
ค่า batch size	16
ค่าสัมประสิทธิ์ reconstruction	0.01

สำหรับตัวเลือก loss function ตามที่ได้กล่าวไปใน 2.1.5 นอกจากการใช้ binary cross entropy ที่เป็นมาตราฐานแล้วทางผู้วิจัยได้บวก reconstruction loss เข้าไปโดยในทางปฏิบัตินั้นคือการบวก mean absolute error loss function หรือ mean square error loss function เข้าไปโดยในที่นี้จะเลือกใช้ mean absolute error loss function ตาม (Isola et al., 2017) ในส่วนของค่า learning rate เราได้ใช้เทคนิคที่ชื่อว่า two-time scale update rule ที่นำเสนอโดย (Heusel et al., 2017) การใช้ค่า learning rate ที่แตกต่างกันระหว่าง discriminator และ generator นั้นจะทำให้การ training มีเสถียรภาพมากขึ้น นอกจากนี้โดยทั่วไปแล้วการสร้างสถาปัตยกรรมของ generator นั้นจะใช้เลเยอร์ convolution transpose ทางผู้วิจัยได้ใช้เทคนิคเลเยอร์ convolution คู่กับเลเยอร์ nearest neighbor upsampling (Odena et al., 2016) เนื่องจากมีการศึกษาว่าการใช้เลเยอร์ convolution transpose นั้นจะทำให้แพทเทิร์นลายตารางหมากรุก (checker board artifact) เกิดขึ้นทำให้ภาพที่สังเคราะห์มาคุณภาพไม่ดีการใช้เทคนิคเลเยอร์ convolution คู่กับเลเยอร์ nearest neighbor upsampling นั้นจะช่วยลดปัญหานี้ ท้ายที่สุดเพื่อทำให้ตัวแบบ GANs เป็นแบบ conditional เราจึงใช้เลเยอร์ embedding เพื่อเปลี่ยนให้ป้ายกำกับของแต่ละภาพกลายเป็นเมทริกซ์และนำไปซ้อนเข้ากับภาพเอกซเรย์ปอดปกติและภาพเอกซเรย์ปอดบวม

3.4 การทดลองการโจมตี

สำหรับรูปแบบการโจมตีที่ผู้วิจัยนำมาทดลองมีชื่อว่า poisonous label attack โดยนำเสนอโดย (Liu et al., 2021) โดยใช้ GANs ในการสร้างภาพสังเคราะห์ขึ้นมาจากนั้นใช้เทคนิคเปลี่ยนป้ายกำกับของภาพนั้นให้เป็นป้ายกำกับอื่นที่ไม่ถูกต้อง poisonous label attack นั้นเป็นการโจมตีชนิด block box นั้นหมายถึงผู้โจมตีสามารภมีความรู้น้อยนิดเกี่ยวกับเป้าหมายก็สามารถทำการโจมตีได้และเป็นการโจมตีประเภทที่ลดประสิทธิภาพของ model โดยรวม โดยดั้งเดิมวิธีการโจมตีนี้จะใช้โจมตีแบบเฉพาะเจาะจงกับข้อมูลที่อยู่ในคลาสหนึ่งๆแต่สามารถปรับเปลี่ยนให้โจมตีแบบไม่เลือกคลาสได้ อีกทั้งการโจมตียังเป็นการโจมตีที่ต้องการจะให้ข้อมูลเป้าหมายถูกทำนายเป็นคลาสอื่นใดก็ได้ที่ไม่ใช่คลาสที่ถูกต้อง

ในส่วนของการได้มาซึ่ง GANs ได้นำเสนอไปแล้วใน 3.3 อีกส่วนคือเทคนิคการติดป้ายกำกับผิดโดย (Liu et al., 2021) โดยนำเสนอสองวิธีคือ symmetric poisoning vector และ asymmetric poisoning vector ทางผู้วิจัยเลือกที่จะใช้ asymmetric poisoning vector เนื่องจาก symmetric poisoning vector นั้นเพียงคือเปลี่ยนป้ายกำกับของภาพไปเป็นป้ายอื่นแบบสุ่มด้วยความน่าจะเป็นที่เท่ากัน (uniform distribution) และรายงานโดย (Liu et al., 2021) ว่ามีประสิทธิภาพที่ไม่ดีแต่ asymmetric poisoning vector นั้นจะเปลี่ยนป้ายกำกับไปเป็นป้ายกำกับอื่นโดยเฉพาะเจาะจงที่ไม่ใช่ป้ายกำกับที่ถูกต้อง (two-point distribution) และมีประสิทธิภาพที่ในการโจมตีที่ดีกว่า

ตามที่ได้กล่าวไปใน 1.3 ขอบเขตการศึกษาเนื่องจาก model เป้าหมายของเราเป็นตัวจำแนกแบบไบนารีการใช้ asymmetric poisoning vector นั้นคือการเปลี่ยนป้ายกำกับของภาพเอกซเรย์ปอดปกติสังเคราะห์ไปเป็นป้ายกำกับของภาพเอกซเรย์ปอดบวมหรือการเปลี่ยนนป้ายกำกับของภาพเอกซเรย์ปอดบวมสังเคราะห์ไปเป็นป้ายกำกับของภาพเอกซเรย์ปอดปกติสังเคราะห์ หากเปลี่ยนแค่กรณีใดกรณีหนึ่งการโจมตีจะเป็นแบบเฉพาะเจาะจง (target attack) แต่สามารถปรับเปลี่ยนให้เป็นการโจมตีแบบไม่เลือก(indiscriminate) ได้โดยการเปลี่ยนทั้งสองกรณี ทางผู้วิจัยเลือกที่ปรับเปลี่ยนการโจมตีเป็นแบบไม่เลือกเนื่องจากการเปลี่ยนป้ายกำกับเพียงแค่คลาสใดคลาสหนึ่งจะทำให้การกระจายตัวของคลาสที่ model เคยเรียนรู้มาเปลี่ยนแปลง

ตารางที่ 7 ตารางแสดงอัลกอริทึมการทดลองการโจมตี

อัลกอริทึม 2 การทดลอง poisonous label attack การโจมตีแบบไม่เลือก

ตัวแปร : M target trained model, G trained generator, z random noise $\sim N(0,1)$

C_1 label of normal lung x-ray image C_2 label of penumonia lung x-ray image

X_1 generated normal lung x-ray image X_2 generated penumonia lung x-ray image

Xp all generated lung x-ray image y_p poisonous labels

X_{test} test set data y_{test} labels of test set

1. **for** each model M
2. Initialize $ACC \leftarrow [\,], SENS \leftarrow [\,], SPEC \leftarrow [\,]$
3. **do** 3 times
4. $M \leftarrow get\ model$
5. Initialize $acc \leftarrow [\,], sens \leftarrow [\,], spec \leftarrow [\,]$
6. While count < maximum poisons
7. $X_1 \leftarrow G(z, C_1), X_2 \leftarrow G(z, C_2)$
8. $X_p \leftarrow [X_1, X_2], y_p \leftarrow [C_2, C_1]$
9. train $M(X_p, y_p)$
10. $predictions \leftarrow M(X_{test})$
11. $accuracy \leftarrow evaluate(predictions, y_{test})$
12. $sensitivity \leftarrow evaluate(predictions, y_{test})$
13. $specificity \leftarrow evaluate(predictions, y_{test})$
14. Append $accuracy$ to acc
15. Append $sensitivity$ to $sens$
16. Append $specigicity$ to $spec$
17. append $acc, sens, spec$ to $ACC, SENS, SPEC$
18. end

3.5 การวัดความคงทนของ model

Security evaluation curve ที่ได้อธิบายไปใน 2.1.1 จะนำมาใช้เพื่อประเมินความคงทนของ deep learning model ที่มีสถาปัตยกรรมที่แตกต่างกันภายใต้การโจมตีที่มีพลังในการโจมตีหรือในที่นี้คือจำนวน poison ที่ใส่เข้าไปในระบบเพิ่มขึ้นโดยตัววัดประสิทธิภาพที่เลือกใช้คือ

1. ค่าความถูกต้อง (accuracy) เป็นสัดส่วนของการทำนายที่ถูกต้อง นั้นคือผลบวกจริง *(True Positive)* และผลลบจริง (True Negative) โดยสามารถเขียนเป็นสมการได้ดังนี้

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

โดยที่

TP คือ ผลบวกจริง

FP คือ ผลลบจริง

FP คือ ผลบวกลวง

FN คือ ผลลบลวง

2. ค่าความไวหรือค่าความระลึก (sensitivity, recall) คือ ค่าวัดประสิทธิภาพของ model ที่สามารถระบุผลบวกจริงได้อย่างถูกต้อง เช่น สามารถระบุได้อย่างถูกต้องว่ามีคนเป็นผู้ป่วยปอดบวมกี่คน

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

3. ค่าความเฉพาะเจาะจง (specificity) คือค่าวัดประสิทธิภาพของ model ที่สามารถระบุผลลบจริงได้ถูกต้อง เช่น สามารถระบุได้อย่างถูกต้องว่ามีคนที่มีปอดปกติกี่คนจากกลุ่มคนทั้งหมด

$$specificity = \frac{TN}{TN + FP}$$

ทำการเก็บค่าตัววัดเหล่านี้เมื่อทำการเพิ่มระดับการโจมตีขึ้นแล้วนำมาสร้างเป็นกราฟ security evaluation curve เพื่อวิเคราะห์ความคงทนของ deep learning model

บทที่ 4
การทดลองและผลการทดลอง

4.1 ระบบและเฟรมเวิร์กที่ใช้ในการทดลอง

ในการทดลองนี้จะใช้ Google Colab pro plus ในการทำการทดลองและใช้เฟรมเวิร์กคือ tensorflow สำหรับการพัฒนา deep learning model เวอร์ชั่น 2.10.0

4.2 ประสิทธิภาพของ model เป้าหมาย

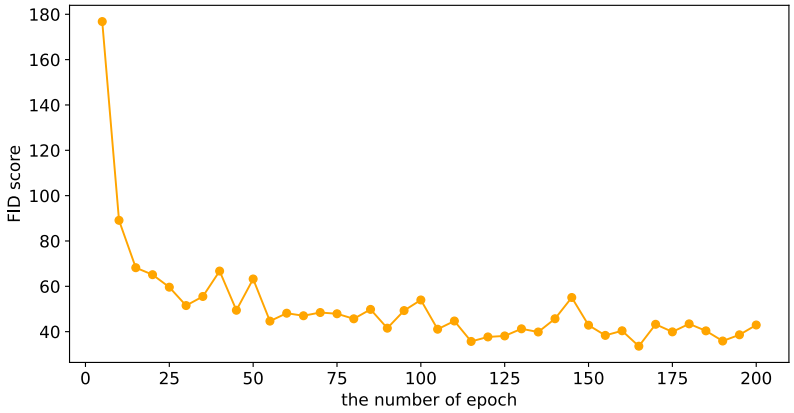
ทำการพัฒนา deep learning model ที่ได้กล่าวไปใน 3.2 โดยใช้การตั้งค่าตามที่กล่าวไป เมื่อการพัฒนาเสร็จสิ้นนำ model ที่เลือกในแต่ละสถาปัตยกรรมมาประเมินประสิทธิภาพกับชุดข้อมูล test set โดยได้ผลดังนี้

ตารางที่ 8 ตารางแสดงประสิทธิภาพของ model เป้าหมายที่ชุดข้อมูล test set

สถาปัตยกรรม	ค่า accuracy	ค่า sensitivity	ค่า specificity
VGG16	0.9463	0.9387	0.9542
ResNet50v2	0.9313	0.9387	0.9234
MobileNetv2	0.9360	0.9316	0.9406
Inceptionv3	0.9367	0.9348	0.9388
ConvNext-Tiny	0.9369	0.9243	0.9489

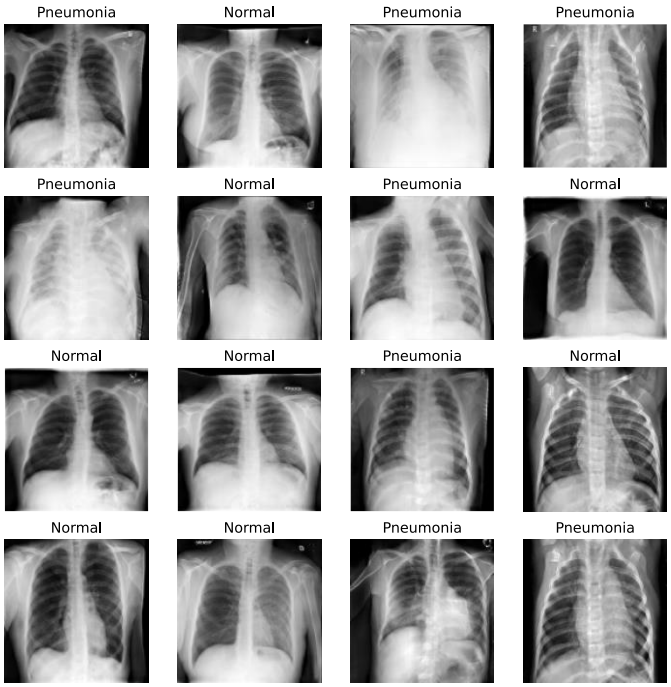
4.3 การประเมิน GANs

ทำการพัฒนา GANs ตามลายละเอียดใน 3.3 โดยในการ training นั้นทำการทำซ้ำทั้งหมด 200 รอบ และทุกๆ 5 รอบจะคำนวณค่า FID 3 ค่าเพื่อมาหาค่าเฉลี่ยเพื่อดูคุณภาพของ model เมื่อการ training ดำเนินไป สำหรับรูปแบบของ FID ที่ใช้นั้นผู้วิจัยเลือกใช้ในรูปแบบของ (Parmar et al., 2022) ซึ่งเป็นการคำนวณที่นำเรื่องการปรับขนาดของภาพเข้ามาพิจารณด้วยและได้ผลค่า FID เฉลี่ยตลอดการ training ตามรูปด้านล่าง



ภาพที่ 12 กราฟแสดงค่า FID เฉลี่ยทุกๆ 5 epoch

จากรูปด้านบนแสดงค่า FID เฉลี่ยจะได้ว่าจุดที่มีค่า FID เฉลี่ยต่ำสุดนั้นคือ epoch ที่ 165 และได้ค่า FID เฉลี่ยอยู่ที่ 33.58 ดังนั้นผู้วิจัยจึงนำ generator ที่ได้จาก GANs ที่ epoch ที่ 165 นั้นไปใช้ทำการทดลองต่อโดยตัวอย่างภาพที่ได้จาก generator ที่ epoch ที่ 165 เป็นดังนี้

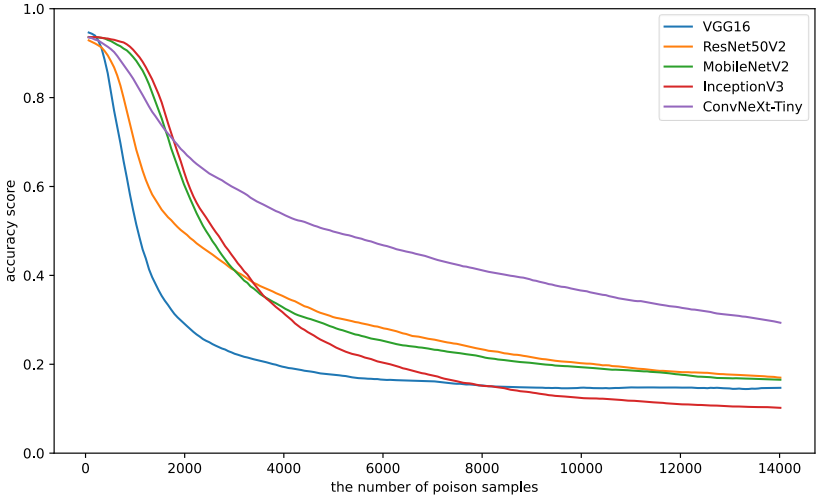


ภาพที่ 13 ตัวอย่างภาพเอกซเรย์สังเคราะห์

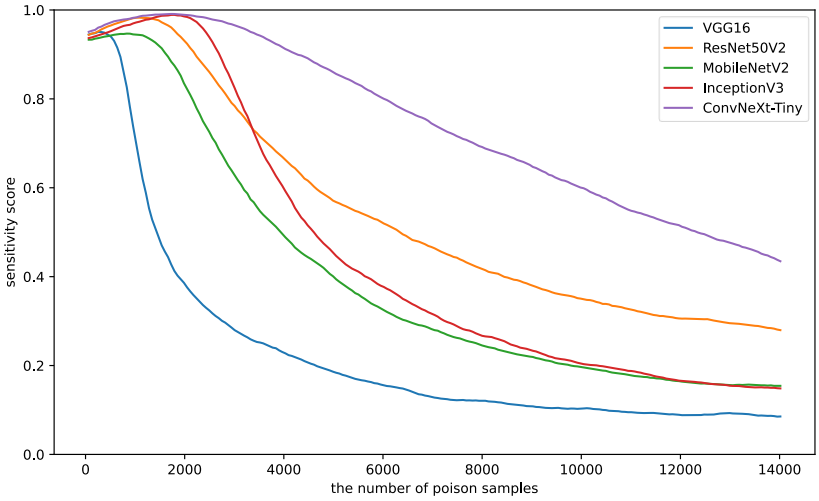
4.4 ผลการทดลองการโจมตี

หลังจากที่ทำการทดลอง 4.2 และ 4.3 เสร็จแล้วนั้น จึงใช้อัลกอริทึมในตารางที่ 9 มาทำการทดลองการโจมตี โดยผู้วิจัยตั้งต้นจาก model ที่ถูกสอนมาแล้วและมีประสิทธิภาพตามตาราง 10 ทำการอัพเดท 5 model นี้ด้วย poison ที่สร้างมาจาก conditional DCGAN และป้ายกำกับที่ผิด เริ่มต้นด้วยการสร้างภาพเอกซเรย์ปอดปกติสังเคราะห์และภาพเอกซเรย์ปอดบวมสังเคราะห์อย่างละ 32 ภาพ รวมแล้ว 64 ภาพเท่ากับค่า batch size ทำการสลับป้ายกำกับของภาพตามบรรทัดที่ 8 ในตาราง 9 หลังจากนั้นจึงนำไปอัพเดท model โดยใช้ค่า learning rate เท่ากันตอน training นั่นคือ 0.0001 จากนั้นตั้งค่า maximum poison ไว้ที่ 14000 เนื่องจากเป็นค่าประมาณจำนวนภาพในชุดข้อมูล training set ทั้งหมดหรือพูดอีกอย่างคือเราทำการอัพเดท poison 1 epoch

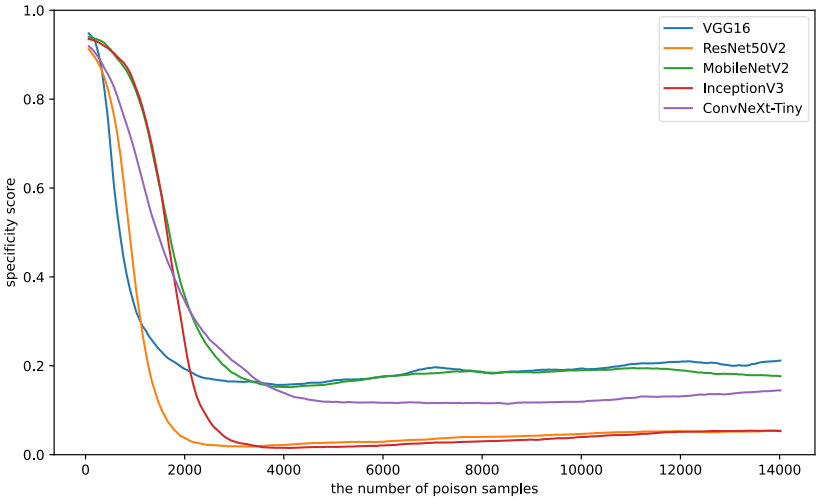
สำหรับการอัพเดท model ในแต่ละสถาปัตยกรรมนั้นผู้วิจัยทำซ้ำทั้งหมด 3 ครั้ง จากนั้นจึงนำค่าตัววัดที่ลดลงมาหาค่าเฉลี่ยและได้ผลตามรูปด้านล่าง



ภาพที่ 14 กราฟแสดงค่าตัววัด accuracy ที่เมื่อเพิ่ม poison



ภาพที่ 15 กราฟแสดงค่าตัววัด sensitivity ที่เมื่อเพิ่ม poison



ภาพที่ 16 กราฟแสดงค่าตัววัด specificity ที่เมื่อเพิ่ม poison

ภาพที่ 12 – 14 จะเห็นว่าสำหรับตัววัด accuracy และ sensitivity นั้น ConvNext-Tiny จะเป็นสถาปัตยกรรมที่มีความคงทนต่อการโจมตีมากที่สุดตามมาด้วย ResNet50v2 และ VGG16 นั้นเป็นสถาปัตยกรรมที่อ่อนแอต่อการโจมตีที่สุด ภาพที่ 14 แสดงให้เห็นว่า model ทุกสถาปัตยกรรมนั้นมีปัญหาหารทำนายภาพเอกซเรย์ปอดปกติเห็นได้จากการที่เส้นกราฟดิ่งลงอย่างรวดเร็วแม้ว่าระดับการโจมตีจะยังน้อย ในภาพที่ 13 จะเห็นว่า model แสดงพฤติกรรมที่น่าสนใจอย่างหนึ่งคือในช่วงเริ่มต้นของการโจมตีค่าตัดวัด sensitivity นั้นมีค่าสูงขึ้นในทุก model แต่ทว่าเมื่อการทดลองดำเนินไปค่าตัววัดก็ถูกลงไปตามระดับการโจมตี

4.5 อภิปรายผลการทดลอง

สถาปัตยกรรมของ ConvNext-Tiny นั้นตามที่ได้อธิบายไว้ในงานวิจัย (Liu et al., 2022) มีลักษณะที่คล้ายคลึงหลายอย่างกับสถาปัตยกรรมแบบ ResNet แต่ทว่าผู้เขียนได้ใช้เทคนิคต่างๆจาก Swin Transformer เข้ามาใส่ในสถาปัตยกรรม ResNet อาจจะเป็นเหตุผลที่ ResNet50v2 นั้นมีความคงทนรองลงมาจาก ConvNext-Tiny ในการทดลองและอาจกล่าวได้ว่าเทคโนโลยีที่มาจาก Swin Transformer นั้นมีส่วนช่วยสนับสนุนความคงทนต่อการโจมตีมากกว่าสถาปัตยกรรมแบบเก่าอย่าง VGG16 นอกจากนี้ ConvNext-Tiny และ ResNet50v2 นั้นมีลักษณะสำคัญร่วมกันคือทั้งคือใช้สถาปัตยกรรมแบบ residual connection ซึ่งก็อาจจะมีส่วนช่วยในการลดผลกระทบจาการโจมตี จากภาพที่ 14 จะเห็นได้ว่า poisonous label attack นั้นสามารถที่จะเพิ่มอัตราผลบวกลวงให้แก่ model ได้ไม่ว่าจะเป็นสถาปัตยกรรมแบบไหนซึ่งเป็นเหตุให้ค่าตัววัด specificity นั้นลดลงอย่างรวดเร็วแสดงให้เห็นว่า poisonous label attack นั้นประสบความสำเร็จในการโจมตี deep learning model ที่มีสถาปัตยกรรมล้ำสมัยและยังเป็นการเน้นย้ำถึงความอ่อนแอของ deep learning model ต่อการโจมตีแบบ poisoning attack แต่ถึงอย่างไรก็ตามการโจมตีนี้ไม่ได้ส่งผลให้ผลลบลวงเพิ่มขึ้นในอัตราที่รวดเร็วเหมือนกับที่ส่งผลให้กับผลบวกลวงแต่การโจมตีนี้ก็ยังสามารถลดค่าตัววัด sensitivity ได้อยู่ดี

ถึงจุดนี้ผู้วิจัยได้แสดงให้เห็นว่า deep learning model สถาปัตยกรรมล้ำสมัยสามารถถูกลดทอนประสิทธิภาพได้วิธีการที่เสนอในวิทยานิพนธ์ฉบับนี้นั้นแตกต่างจากงานวิจัยที่ผ่านโดยที่ผู้วิจัยเลือกใช้สถาปัตยกรรมที่ล้ำสมัยและใช้ชุดข้อมูลในโลกความจริงในงานทางการแพทย์ผลการทดลองของงานศึกษานี้ได้เน้นย้ำในเรื่องของการต้องการสถาปัตยกรรมที่มีความคงทนต่อการโจมตีในงานที่ต้องคำนึงถึงชีวิตอย่างงานทางการแพทย์ นักพัฒนาสามารถใช้สถาปัตยกรรมที่ถูกแนะนำในงานศึกษานี้เพื่อต่อยอดสร้าง deep learning model ที่ปลอดภัยต่อไปได้

บทที่ 5
สรุปผลการทดลอง

วิทยานิพนธ์ฉบับนี้ผู้วิจัยได้ทำการทดลองโจมตี deep learning model ด้วยวิธี poisonous label attack และแสดงให้เห็นว่า model ที่มีสถาปัตยกรรมล้ำสมัยที่ถูกพัฒนามาเพื่อจำแนกภาพทางการแพทย์แบบไบนารีนั้นสามารถถูกลดทอนประสิทธิภาพได้ ผู้วิจัยใช้ conditional DCGAN ที่มีค่า FID เฉลี่ย 33.58 ในการสร้างภาพเอกซเรย์ปอดสังเคราะห์ขึ้นมาและสลับป้ายกำกับของภาพสังเคราะห์ จากนั้นจึงอัพเดท model ที่ถูกพัฒนามาแล้วด้วย poison และสังเกตบันทึกถึงประสิทธิที่ลดลง ผู้วิจับค้นพบว่าสถาปัตยกรรม ConvNext นั้นมีความคงทนต่อการโจมตีมากที่สุดตามมาด้วย ResNet50v2 นี้เองอาจจะสามารถบอกได้ว่าเทคโนโลยีที่มาจาก Transformer และการใช้ residual connection นั้นมีส่วนช่วยสนับสนุนความคงทนของ model และสถาปัตยกรรมแบบพื้นฐานอย่าง VGG16 ที่นำเลเยอร์ convolution มาซ้อนกันนั้นควรที่จะหลีกเลี่ยง

ในอนาคตแนวคิดต่อยอดคือการทำการทดลองกับการโจมตี poisoning attack แบบอื่นๆกับ model สถาปัตยกรรมล้ำสมัยเฉพาะทางการแพทย์และทั่วไปยกตัวอย่างเช่น Swin Transfomer, Vision Transformer และ model อื่นๆที่มีพื้นฐานมาจากกลไก attention นอกจากนี้การทดลองเพื่อดูความสามารถในการแปรผลของ model (interpertability) ภายใต้การโจมตีก็ควรค่าที่จะสำรวจ

บรรณานุกรม

Asgari Taghanaki, S., Das, A., & Hamarneh, G. (2018). Vulnerability analysis of chest X-ray image classification against adversarial attacks. In *Understanding and interpreting machine learning in medical image computing applications* (pp. 87-94). Springer.

Bhagat, V., & Bhaumik, S. (2019). Data augmentation using generative adversarial networks for pneumonia classification in chest Xrays. 2019 Fifth International Conference on Image Information Processing (ICIIP),

Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84, 317-331.

Finlayson, S. G., Chung, H. W., Kohane, I. S., & Beam, A. L. (2018). Adversarial attacks against medical deep learning systems. *arXiv preprint arXiv:*1804.05296.

Goodfellow, I. (2016). Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:*1701.00160.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Networks. arXiv:1406.2661. Retrieved June 01, 2014, from https://ui.adsabs.harvard.edu/abs/2014arXiv1406.2661G

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:*1412.6572.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. Proceedings of the IEEE conference on computer vision and pattern recognition,

Kasichainula, K., Mansourifar, H., & Shi, W. (2021). Poisoning Attacks via Generative Adversarial Text to Image Synthesis. 2021 51st Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W),

Kim, D., Joo, J., & Kim, S. C. (2022). Fake Data Generation for Medical Image

Augmentation using GANs. 2022 International Conference on Artificial Intelligence in Information and Communication (ICAIIC),

Kora Venu, S., & Ravula, S. (2020). Evaluation of deep convolutional generative adversarial networks for data augmentation of chest x-ray images. *Future Internet*, 13(1), 8.

Li, Y., Xiao, N., & Ouyang, W. (2019). Improved generative adversarial networks with reconstruction loss. *Neurocomputing*, 323, 363-372.

Liu, H., Li, D., & Li, Y. (2021). Poisonous Label Attack: Black-Box Data Poisoning Attack with Enhanced Conditional DCGAN. *Neural Processing Letters*, 53(6), 4117-4142.

Liu, X., Xie, L., Wang, Y., Zou, J., Xiong, J., Ying, Z., & Vasilakos, A. V. (2020). Privacy and security issues in deep learning: A survey. *IEEE access*, 9, 4566-4593.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv*:1411.1784.

Mozaffari-Kermani, M., Sur-Kolay, S., Raghunathan, A., & Jha, N. K. (2015). Systematic Poisoning Attacks on and Defenses for Machine Learning in Healthcare. *IEEE J Biomed Health Inform*, 19(6), 1893-1905. https://doi.org/10.1109/JBHI.2014.2344095

Muñoz-González, L., Biggio, B., Demontis, A., Paudice, A., Wongrassamee, V., Lupu, E. C., & Roli, F. (2017). Towards poisoning of deep learning algorithms with back-gradient optimization. Proceedings of the 10th ACM workshop on artificial intelligence and security,

Odena, A., Dumoulin, V., & Olah, C. (2016). Deconvolution and checkerboard artifacts. *Distill*, 1(10), e3.

Parmar, G., Zhang, R., & Zhu, J.-Y. (2022). On aliased resizing and surprising subtleties in gan evaluation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Qayyum, A., Qadir, J., Bilal, M., & Al-Fuqaha, A. (2020). Secure and robust machine learning for healthcare: A survey. *IEEE Reviews in Biomedical Engineering*, 14, 156-180.

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:*1511.06434.

Shafahi, A., Huang, W. R., Najibi, M., Suciu, O., Studer, C., Dumitras, T., & Goldstein, T. (2018). Poison frogs! targeted clean-label poisoning attacks on neural networks. *Advances in neural information processing systems*, 31.

Shi, Y., Sagduyu, Y. E., Davaslioglu, K., & Li, J. H. (2018). Generative adversarial networks for black-box API attacks with limited training data. 2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT),

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:*1312.6199.

Tahir, A. M., Chowdhury, M. E., Khandakar, A., Rahman, T., Qiblawey, Y., Khurshid, U., Kiranyaz, S., Ibtehaz, N., Rahman, M. S., & Al-Maadeed, S. (2021). COVID-19 infection localization and severity grading from chest X-ray images. *Computers in biology and medicine*, 139, 105002.

Yang, C., Wu, Q., Li, H., & Chen, Y. (2017). Generative poisoning attack method against neural networks. *arXiv preprint arXiv:*1703.01340.

Zhang, Q., Wang, H., Lu, H., Won, D., & Yoon, S. W. (2018). Medical image synthesis with generative adversarial networks for tissue recognition. 2018 IEEE International Conference on Healthcare Informatics (ICHI),

Zhou, S. K., Greenspan, H., Davatzikos, C., Duncan, J. S., Van Ginneken, B., Madabhushi, A., Prince, J. L., Rueckert, D., & Summers, R. M. (2021). A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, 109(5), 820-838.

ประวัติผู้เขียน

ชื่อ-สกุล	Pakpoom Singkorapoom
วัน เดือน ปี เกิด	02 Nov 1996
สถานที่เกิด	Bangkok
วุฒิการศึกษา	Department of Mechanical Engineering, Kasetsart University. Department of Statistics, Chulalongkorn University
ที่อยู่ปัจจุบัน	251 Supalai ville soi 14/3, Rattanathibeth Road, Moung, Nonthaburi, Thaialnd 11000