

การจำแนกกลุ่มของสัญญาณโจมตีแบบจำลองการเรียนรู้เชิงลึก

Clustering Analysis of Adversarial Perturbations on Deep Learning Models

ศิริกร ลำไย*, วัชรพัฐ เมตตานันท์†, และ จิตรัทศน์ ผักเจริญผล‡
ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์
Email: *sirakorn.l@ku.th, †vacharapat@eng.src.ku.ac.th, ‡jtf@ku.ac.th

บทคัดย่อ

บทความวิชาการนี้กล่าวถึงการฝึกสอนแบบจำลองอย่างง่าย เพื่อค้นหารูปแบบของสัญญาณรบกวนที่สามารถโจมตีชุดข้อมูลรับเข้า ให้ได้ผลลัพธ์ของแบบจำลองที่ผิดเพี้ยนไปได้ โดยพิจารณาการโจมตีบนโครงข่ายประสาทเทียมแบบเชื่อมถึงกันทั่ว ซึ่งถูกฝึกสอนด้วยชุดข้อมูล MNIST ซึ่งมีชุดข้อมูลสำหรับฝึกสอนจำนวน 60,000 จุด หลังจากการฝึกสอนโครงข่ายประสาทเทียม นำข้อมูลทดสอบจำนวน 10,000 จุด มาทำการหาสัญญาณโจมตีความยาวเท่าจำนวนจุดทดสอบ และพยายามทำการเรียนรู้จัดหมวดหมู่ ผลลัพธ์ที่ได้คือคะแนนเอฟ-1 ของบางชั้นข้อมูลลดลงไปไม่มากเมื่อเทียบกับชั้นข้อมูลอื่น และผลการจัดหมวดหมู่แสดงให้เห็นถึงความสัมพันธ์บางอย่างระหว่างข้อมูลเป้าหมายตั้งต้นและคำตอบที่แบบจำลองเห็น ซึ่งอาจนำไปสู่ประเด็นใหม่ในการศึกษาการโจมตีแบบจำลองต่อไป

คำสำคัญ: ปัญญาประดิษฐ์, จักรกลเรียนรู้, การเรียนรู้เชิงลึก, การโจมตีการเรียนรู้

Abstract

This work present the process of training a simple deep learning model in order to find the set of perturbations that attacks the model, therefore giving out the incorrect responses. We train a fully connected deep learning model to recognise MNIST handwriting database with 60,000 training points, and find 10,000 perturbations, each to attack 10,000 testing points. The results shows a decrease in model's F-1 score after the perturbations were mixed along with the testing points. However, the performances of all classes does not equally degrade. Moreover, the clustering analysis shows some intracluster relationship between the original labels of the testing points, and the labels classified by the model. This work could lightens new scopes and perspectives on

the behaviour of deep learning models after attempts on adversarial attacks.

Keywords: Artificial Intelligence, Machine Learning, Deep Learning, Adversarial Attack

1. ความสำคัญและที่มา

แบบจำลองจักรกลเรียนรู้ (machine learning models) นั้นถูกใช้อย่างกว้างขวางในปัจจุบัน อย่างไรก็ตามแบบจำลองใดๆ นั้นอาจมีความผิดพลาดต่อการทำการโจมตีประสงค์ร้าย (adversarial attacks) เพื่อจงใจให้ผลลัพธ์ที่แบบจำลองนั้นคาดเดามีความผิดพลาดจากผลลัพธ์ที่ควรจะเป็น [1]

ในการเรียนรู้เชิงตัวแปรเสริม (parameter-based learning) นั้น ตัวแปรเสริม (parameters) ค่าน้ำหนัก (weights) บนแบบจำลองการเรียนรู้เชิงลึก (deep learning models) เป็นตัวกำหนดความฉลาดของแบบจำลอง อาจมีตัวแปรเสริมบางชุดที่ทำให้แบบจำลองมีช่องโหว่ต่อการโจมตีประสงค์ร้าย การโจมตีนั้นอาจเกิดจากการเพิ่มสัญญาณรบกวนซึ่งผ่านการคำนวณ (calculated artefacts) เข้าสู่ข้อมูลรับเข้า (inputs) ซึ่งทำให้ความผิดพลาดของแบบจำลองในการพยากรณ์คำตอบนั้นเปลี่ยนไปอย่างชัดเจน

2. ขบวนการกรรมและกระบวนการวิธี

2.1. การโจมตีการเรียนรู้ [2], [3]

เราจะกล่าวถึงแบบจำลองที่ถูกฝึกสอนให้จัดจำแนกข้อมูลชุด X และ Y และพิจารณาข้อมูลรับเข้า (x, y) หนึ่งจุดบนชุดทดสอบจะนิยามข้อมูลโจมตี (adversarial) \tilde{x} ว่า

$$\tilde{x} = x + \eta \quad (1)$$

เมื่อเรียก η ว่าสัญญาณรบกวน (perturbations)

ข้อสังเกตที่เกิดขึ้นคือเราอาจนิยามชุดสัญญาณรบกวนดังกล่าว ว่ามีความเข้มข้น (intensity) ในระดับที่ต่ำกว่าตามนุษย์จะ

มองเห็น กล่าวคือเมื่อเทียบกับชุดข้อมูลรับเข้าแล้ว ช่วง (range) ของสัญญาณรบกวนนั้นน้อยกว่าช่วงของข้อมูลรับเข้าที่เป็นไปได้มาก การนิยามดังกล่าวจะใช้การนิยามเซตของสัญญาณรบกวนที่เป็นไปได้ทั้งหมด (พิจารณาว่ามีค่า η ที่เป็นไปได้หลายค่า และแต่ละค่าโจมตีแบบจำลองได้แตกต่างกันออกไป) ว่า

$$H = \{\eta : \|\eta\|_\infty \leq \epsilon\} \quad (2)$$

เมื่อนิยามให้ตัวดำเนินการอนันต์ (infinity norm) เป็น

$$\|x\|_\infty = \max_i x_i \quad (3)$$

และค่า ϵ เป็นค่าคงที่บ่งบอกความเข้มข้นของสัญญาณมากที่สุดที่ได้รับได้ โดยมากมักมีค่าน้อย

2.2. ฟังก์ชันสูญเสีย และการฝึกสอนแบบจำลองด้วยวิธีก้าวเคลื่อนถอยหลัง

พิจารณาการเรียนรู้แบบจำลอง M จะพบว่าการหาตัวแปรเสริม (parameters) θ ที่ดีที่สุดของ M นั้นทำได้ด้วยการนิยามฟังก์ชันสูญเสีย (loss function) ℓ_i ของจุดฝึกหัด (training point) i ได้ โดยให้ฟังก์ชันสูญเสียเป็นฟังก์ชันที่เปรียบเทียบเป้าหมาย (target) y_i จากชุดฝึกหัด และคำตอบ $\hat{y}_i = M(x_i)$ จากชุดคุณสมบัติ (features) x_i ที่ถูกป้อนเข้าแบบจำลอง

เราอาจนิยามฟังก์ชันสูญเสียอย่างง่ายได้เป็นฟังก์ชันผลของผลต่างกำลังสอง

$$\ell_i = \sum_{i=1}^M (\hat{y}_i - y_i)^2 = \sum_{i=1}^M (M(x_i) - y_i)^2 \quad (4)$$

เมื่อ M เป็นขนาดของเป้าหมาย (target) สังเกตว่ายิ่งค่าของ \hat{y}_i และ y_i ต่างกันมากเท่าใด (มองอีกมุมหนึ่ง คือยิ่งตอบผิดมากเท่าใด) ค่าดังกล่าวก็จะยิ่งเพิ่มขึ้นมากเท่านั้น อย่างไรก็ตาม ในกรณีของการเรียนรู้เชิงการจำแนก (classification) ส่วนมาก มักใช้ฟังก์ชันสูญเสียเป็นฟังก์ชันสูญเสียแบบความวุ่นวายข้ามชั้น (cross entropy loss)

$$\ell_i = - \sum_{c=1}^M y_{o,c} \ln(p_{o,c}) \quad (5)$$

เมื่อ M เป็นจำนวนชั้น (class) ที่เป็นไปได้ y เป็นค่าฐานสองที่บ่งบอกว่าชั้นข้อมูล (class) c เป็นคำตอบที่ถูกต้องสำหรับการคาดเดา (observation) o และ p เป็นค่าความน่าจะเป็นที่การคาดเดา o ตอบว่าเป็นชั้นข้อมูล c

นอกจากนี้เราอาจนิยามผลรวมของฟังก์ชันสูญเสียทั่วทั้งชุด

ฝึกสอน

$$\mathcal{L} = \sum_{i=1}^N \ell_i \quad (6)$$

เป็นผลรวมของฟังก์ชันสูญเสียบนทุกจุดฝึกหัด เมื่อ N เป็นขนาดของชุดฝึกหัด (training set)

อย่างไรก็ดี แม้สมการ 4 และ 5 จะดูเหมือนพิจารณาค่าสูญเสียที่เปลี่ยนไปเมื่อชุดของข้อมูลฝึกหัดเปลี่ยน แต่พึงระวังว่าการนิยามฟังก์ชันสูญเสียดังกล่าว มีขึ้นเพื่อทดสอบว่าค่าตัวแปรเสริม θ ใดๆ ส่งผลให้แบบจำลองให้คำตอบผิดเพี้ยนมากหรือน้อยเพียงใด สังเกตว่าการเปลี่ยนค่า θ จะส่งผลให้ค่าของ \hat{y} และ p ในทั้งสองสมการตามลำดับเปลี่ยนไป และทำให้ความถูกต้องของแบบจำลองเปลี่ยนไปเช่นกัน ดังนั้นเรามักเขียนฟังก์ชันสูญเสียในสมการที่ 6 ใหม่ให้รับค่าตัวแปรเสริม θ เข้ามาเป็น

$$\mathcal{L}(\theta) = \sum_{i=1}^N \ell_i \quad (7)$$

การฝึกสอนแบบจำลองการเรียนรู้เชิงลึกมักใช้วิธีการเกรเดียนต์ดลัน (gradient descent) โดยพิจารณาการปรับแบบจำลองอยู่บนเกรเดียนต์ของฟังก์ชันสูญเสีย

$$\theta' = \theta - l \frac{\partial}{\partial \theta} \mathcal{L}(\theta) \quad (8)$$

เมื่อ l เป็นค่าอัตราการเรียนรู้ (learning rate) โดยปกติมักมีค่าไม่มาก

หากอธิบายโดยคร่าว ขั้นตอนวิธีเกรเดียนต์ดลัน พยายามหาค่าตัวแปรเสริม θ_{OPT} โดยการเริ่มจากการสุ่มตัวแปรเสริม θ แล้วคำนวณเกรเดียนต์ของฟังก์ชันสูญเสีย และค่อยๆ ปรับค่า θ ตามทิศตรงข้ามกับเกรเดียนต์เรื่อยๆ จนกระทั่งถึงจุดที่ฟังก์ชันสูญเสียมีค่าน้อยที่สุด

2.3. การหาสัญญาณรบกวนด้วยวิธีการก้าวเคลื่อนถอยหลัง

เมื่อฝึกสอนแบบจำลองการเรียนรู้เชิงลึกโดยได้ชุดตัวแปรเสริม θ สำหรับแบบจำลอง M ซึ่งต่อไปนี้จะเรียกชุดแบบจำลองและตัวแปรเสริมรวมกันว่า M_θ แล้ว เมื่อให้ชุดข้อมูลรับเข้าและส่งออก (x, y) ใดๆ เราอาจหาสัญญาณรบกวนได้ว่า

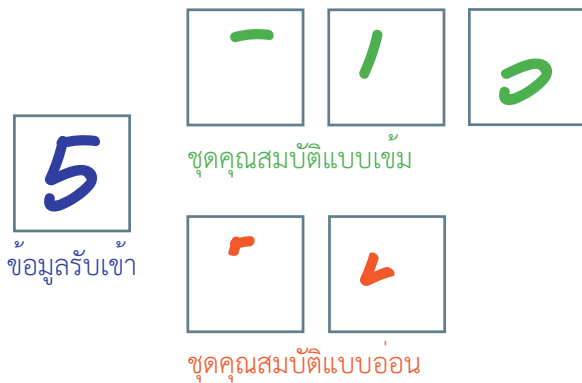
$$\eta' = \eta + l \frac{\partial}{\partial \eta} \mathcal{L}(x) \quad (9)$$

เมื่อ l เป็นค่าอัตราการเรียนรู้ (learning rate) โดยปกติมักมีค่าไม่มาก

จะสังเกตได้ว่าสมการที่ 9 มีลักษณะคล้ายกับสมการที่ 8 เป็นอย่างมาก แตกต่างกันเพียงแต่เครื่องหมายบวกหรือลบ และตัวแปรเทียบสำหรับการทำอนุพันธ์หลายตัวแปร (multivariable derivation) ขอให้สังเกตว่าในขณะที่สมการ 8 พยายามหาค่า θ ที่ทำให้ฟังก์ชันสูญเสีย \mathcal{L} มีค่าต่ำที่สุด สมการที่ 9 กลับพยายามหาสัญญาณรบกวน η ที่ทำให้ฟังก์ชันสูญเสีย \mathcal{L} มีค่ามากที่สุด กล่าวคือตอบผิดมากที่สุด

2.4. คำอธิบายต่อการเกิดขึ้นของสัญญาณรบกวน

มีหลายทฤษฎีพยายามอธิบายการเกิดขึ้นของการโจมตีแบบจำลอง ซึ่งอาจยกตัวอย่างทฤษฎีและคำอธิบายได้ดังนี้



รูปที่ 1. ตัวอย่างชุดคุณสมบัติแบบอ่อน และแบบเข้มที่เป็นไปได้ จากเลข 5

2.4.1. การประพัตต์ตัวเป็นเส้นตรง. LeCun และคณะ [4] ศึกษาผลของการโจมตีที่เกิดจาก \tilde{x} โดยอาจพิจารณาได้จากการคุณสมบัติการเพื่อหาค่าส่งออกจากชุดน้ำหนัก (weights) ของชั้นแบบจำลองการเรียนรู้เชิงลึก (deep learning layers)

$$w^T \tilde{x} = w^T x + w^T \eta \quad (10)$$

คณะวิจัยสังเกตพฤติกรรมว่าสัญญาณรบกวน η กระตุ้นส่วนของชุดน้ำหนักและฟังก์ชันกระตุ้น (activation function) ในแบบจำลองให้ประพัตต์ตัวเยี่ยงฟังก์ชันเส้นตรง (linear functions) ซึ่งการแสดงผลพฤติกรรมดังเส้นตรง (linearity) ในกรณีขยขอบ (edge case) ของข้อมูลรับเข้านั้นก่อให้เกิดความเป็นไปได้ที่แบบจำลองจะถูกโจมตี

เพื่อพิสูจน์ทฤษฎีดังกล่าว Goodfellow และคณะ พิจารณาผลความน่าจะเป็นของคำตอบที่ออกจากแบบจำลองเมื่อปรับค่า ϵ ดังแสดงในสมการที่ 2 และพบว่าความน่าจะเป็นของข้อมูลส่งออก (output) ของแต่ละชั้นข้อมูล (class) มีความสัมพันธ์เชิงเส้นตรงกับค่า ϵ ที่เพิ่มขึ้นเรื่อยๆ

2.4.2. ทฤษฎีชุดคุณสมบัติแบบอ่อนและแบบเข้ม. Ilyas และคณะ [5] ศึกษาโครงสร้างของแบบจำลองเชิงลึก จนนำมาสู่ข้อสรุปว่า “ข้อโหว่ในการโจมตีแบบจำลองเป็นผลโดยตรงจากความอ่อนไหวของแบบจำลองในการวางหลักการบนชุดคุณสมบัติของข้อมูล” (“Adversarial vulnerability is a direct result of our models’ sensitivity to well-generalizing features in the data”)

หากกล่าวให้ละเอียด พิจารณาว่าโครงสร้างของแบบจำลองเชิงลึกสามารถเรียนรู้ชุดคุณสมบัติ (features) ของข้อมูลรับเข้าได้สองแบบ ซึ่งในงานวิจัยเรียกว่าชุดคุณสมบัติแบบอ่อน (weak features) และชุดคุณสมบัติแบบเข้ม (strong features)

- ชุดคุณสมบัติแบบเข้ม คือชุดคุณสมบัติที่มนุษย์มองเห็นโดยทั่วไป กล่าวคือเป็นชุดคุณสมบัติที่มนุษย์สามารถสังเกตทำความเข้าใจ และวางหลักการในการจำแนกได้

เป้าหมาย	ข้อมูลตั้งต้น			ข้อมูลที่ถูกโจมตี		
	พริซิชั่น	รีคอลล์	เอฟ-1	พริซิชั่น	รีคอลล์	เอฟ-1
0	0.93	0.98	0.96	0.88	0.95	0.91
1	0.97	0.97	0.97	0.94	0.91	0.93
2	0.93	0.92	0.93	0.77	0.82	0.80
3	0.91	0.92	0.91	0.74	0.74	0.74
4	0.93	0.93	0.93	0.80	0.82	0.81
5	0.91	0.89	0.90	0.71	0.71	0.71
6	0.93	0.95	0.94	0.89	0.85	0.87
7	0.95	0.91	0.93	0.87	0.80	0.84
8	0.89	0.90	0.89	0.70	0.74	0.72
9	0.92	0.91	0.92	0.80	0.77	0.78

ตารางที่ 1. ตารางแสดงค่าพริซิชั่น (precision) รีคอลล์ (recall) และคะแนน F-1 ของแบบจำลอง ก่อนและหลังการโจมตี

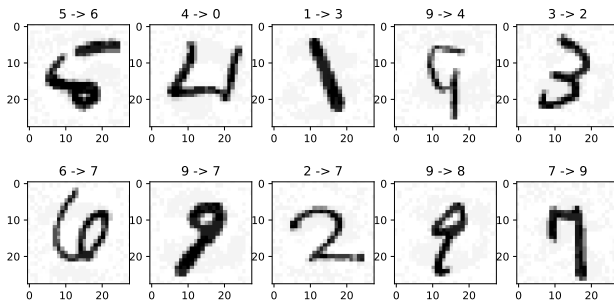
- ชุดคุณสมบัติแบบอ่อน คือชุดคุณสมบัติที่มนุษย์ไม่สามารถมองเห็น หรือมองเห็นแต่ไม่ได้หยิบมาเป็นตัวปัจจัยหลักในการตัดสินใจ และวางหลักการในการจำแนก

จะยกตัวอย่างกรณีการจำแนกเลข 5 เราอาจพิจารณาว่าเลข 5 ดังแสดงในรูปที่ 1 ประกอบขึ้นจากขีดหนึ่งขีดแนวขวาง ขีดหนึ่งขีดแนวตั้ง และส่วนโค้งคล้ายวงกลม เป็นชุดคุณสมบัติที่มนุษย์สังเกตและเข้าใจโดยทั่วไป รวมถึงเป็นคุณสมบัติที่มนุษย์ใช้ในการสังเกตเห็นเส้นที่เชื่อมต่อกันจนประกอบเป็นเลข 5 อย่างไรก็ตาม แบบจำลองการเรียนรู้ใดๆ อาจเห็นมุมมองต่อระหว่างขอบ (ซึ่งอาจสังเกตได้ว่าไม่มีเลขาตัวใดเลยนอกจาก 1 ถึง 9 ยกเว้น 5 ที่มีมุมและขอบดังแสดง) เป็นตัวตัดสินใจในการเรียนรู้เลข 5 อย่างไรก็ตาม พึงสังเกตว่าแบบจำลองอาจจะแม้กระทั่งเลือกสังเกตเห็นพื้นที่ว่างบริเวณที่แตกต่างกันไป และใช้พื้นที่ว่างเหล่านั้นเพื่อสร้างข้อสรุปหรือตัดสินใจว่าเลขที่มองเห็นเป็นเลขใด (ซึ่งการนำมาซึ่ง “ข้อสรุป” จากที่ว่างนั้น ชัดกับวิสัยปกติของมนุษย์ในการสังเกตและมองเห็นอย่างชัดเจน)

2.5. การทดลองหาสัญญาณรบกวนบนชุดข้อมูล MNIST

ในงานชิ้นนี้ ผู้เขียนทำการฝึกสอนแบบจำลองการเรียนรู้เชิงลึกแบบขั้นเชื่อมถึงกันหมด (Fully-connected Deep Learning model) บนฐานข้อมูล MNIST [6] ซึ่งประกอบด้วยชุดฝึกหัดจำนวน 60,000 ข้อมูล มีชุดคุณสมบัติ (features) เป็นรูปภาพลายมือเขียนตัวเลขขนาด 28 พิกเซลแบบขาวดำ และมีชั้นเป้าหมาย (targets) เป็นตัวเลข 1-9 ที่ปรากฏในรูป เมื่อทำการฝึกสอนแบบจำลองเป็นที่เรียบร้อยแล้ว ผู้เขียนหาสัญญาณโจมตี η จำนวน 10,000 จุด บนแต่ละจุดข้อมูลของชุดทดสอบความยาว 10,000 ข้อมูล กล่าวคือสัญญาณโจมตี η_i โจมตีจุดทดสอบ (x_i, y_i)

ข้อมูลคุณสมบัติในทั้งชุดฝึกสอนและชุดทดสอบถูกปรับช่วงข้อมูล (rescale) ให้อยู่ในช่วง $[-1, 1]$ วัดผลด้วยฟังก์ชันสูญเสียแบบความซับซ้อนข้ามชั้น (cross entropy loss) ดังแสดงในสมการที่ 5 ใช้อัตราการเรียนรู้ (learning rate) $l = 0.03$ และฝึกสอนเป็นจำนวน 20 รอบวนซ้ำ (epochs)



รูปที่ 2. ตัวอย่างข้อมูลที่ถูกลบ

2.6. การจำแนกกลุ่มของสัญญาณโจมตี

พิจารณาข้อมูลฝึกหัด (x_i, y_i) ซึ่งถูกสัญญาณโจมตี η_i ทำให้แบบจำลองตอบขึ้นข้อมูลของ x_i ผิดเป็น m_i ผู้เขียนทำการจำแนกกลุ่ม (clustering) ข้อมูลสัญญาณโจมตี η จำนวน 10,000 ตัว โดยใช้จำนวนกลุ่ม (clusters) 10 กลุ่ม และพิจารณาว่าแต่ละกลุ่มของสัญญาณโจมตีนั้นมีเป้าหมาย y_i เดิม และคำตอบที่แบบจำลองตอบออกมาผิด m_i เป็นเท่าใดบ้าง

3. ผลลัพธ์

ตัวอย่างของข้อมูลหลังถูกลบด้วยสัญญาณโจมตี ดังแสดงในรูปที่ 2 พร้อมกับป้ายระบุว่าแบบจำลองเห็นข้อมูลที่ถูกลบเป็นเลขใด

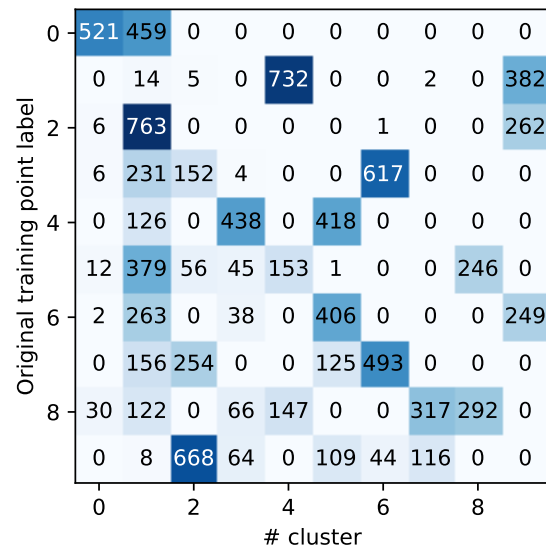
ข้อมูลแสดงอัตราพรีซิชั่น (precision) รีคอลล์ (recall) และคะแนนเอฟ-1 (F-1 score) ของแบบจำลองนั้น ดังแสดงในตารางที่ 1 โดยแยกเป็นกรณีคะแนนของชุดทดสอบก่อนการโจมตี และชุดทดสอบหลังโจมตีด้วยความเข้มสัญญาณ $\epsilon = 0.05$ (ดังแสดงในสมการที่ 2)

อาจพิจารณาได้ว่า คะแนนเอฟ-1 ของบางชั้นข้อมูล (เช่น เลข 1) ลดลงไม่มากเมื่อเทียบกับชั้นข้อมูลอื่น (เช่น เลข 2 และเลข 5) การพิจารณาลักษณะนี้อาจนำมาสู่สมมติฐานว่าเราไม่สามารถโจมตีเป้าหมายของแบบจำลองการเรียนรู้ซึ่งลึกแต่ละแบบได้ด้วยความง่ายเท่ากัน อย่างไรก็ตาม สมมติฐานดังกล่าวต้องการการทดลองอีกเป็นจำนวนมากก่อนจะสามารถสรุปได้ว่าจริงหรือไม่จริง

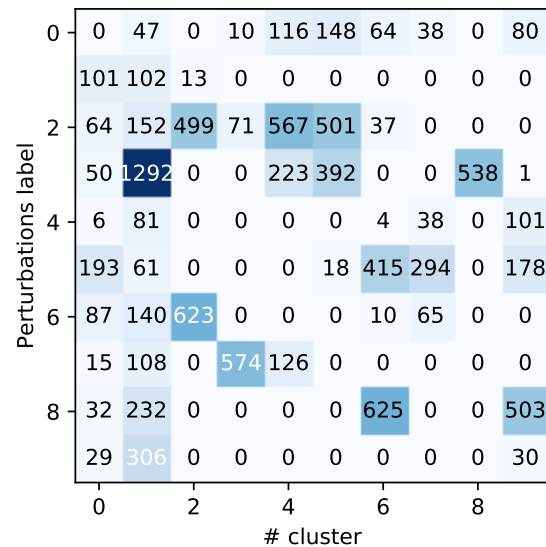
4. อภิปรายผล

4.1. ความแม่นยำของแบบจำลองต่อสัญญาณรบกวน

เมื่อพิจารณาคะแนนเอฟ-1 ซึ่งเป็นค่าเฉลี่ยฮาร์โมนิก (harmonic mean) ของพรีซิชั่น (precision) และรีคอลล์ (recall) อาจพิจารณาได้ว่า คะแนนเอฟ-1 ของบางชั้นข้อมูล (เช่น เลข 1) ลดลงไม่มากเมื่อเทียบกับชั้นข้อมูลอื่น (เช่น เลข 2 และเลข 5) การพิจารณาลักษณะนี้อาจนำมาสู่สมมติฐานว่าเราไม่สามารถโจมตีเป้าหมายของแบบจำลองการเรียนรู้ซึ่งลึกแต่ละแบบได้ด้วยความง่ายเท่ากัน อย่างไรก็ตาม



รูปที่ 3. เมทริกซ์แสดงความฉงน ระหว่างหมายเลขของกลุ่มคลัสเตอร์ และเป้าหมายตั้งต้นของข้อมูล



รูปที่ 4. เมทริกซ์แสดงความฉงน ระหว่างหมายเลขของกลุ่มคลัสเตอร์ และคำตอบที่แบบจำลองเห็น

สมมติฐานดังกล่าวต้องการการทดลองอีกเป็นจำนวนมากก่อนจะสามารถสรุปได้ว่าจริงหรือไม่จริง

4.2. ความแม่นยำของแบบจำลองต่อสัญญาณรบกวน

ภาพที่ 3 แสดงผลการจัดกลุ่ม (clustering) ของสัญญาณรบกวน เทียบกับเป้าหมาย (target) ตั้งต้นของข้อมูล พบว่าในคลัสเตอร์ (cluster) ที่ 1, 2, และ 4 มีสัญญาณรบกวนที่พยายามรบกวนข้อมูลเลข

2, 9, และ 1 มากเป็นพิเศษ และในคลัสเตอร์ที่ 6 มีสัญญาณรบกวนที่พยายามรบกวนข้อมูลเลข 3 และ 7 มากเป็นพิเศษ

ภาพที่ 4 แสดงผลการจัดกลุ่ม (clustering) ของสัญญาณรบกวน เทียบกับคำตอบของเป้าหมายจากแบบจำลอง พบว่าในคลัสเตอร์ (cluster) ที่ 1, 2, และ 4 แบบจำลองมองเห็นสัญญาณรบกวนเป็นเลข 3, 6, และ 2 มากเป็นพิเศษ และในคลัสเตอร์ที่ 6 มีสัญญาณรบกวนที่พยายามรบกวนข้อมูลเลข 5 และ 8 มากเป็นพิเศษ

แม้ผู้เขียนจะยังไม่สามารถสรุปได้แน่ชัดว่าการทำการจัดหมวดหมู่ดังกล่าวแสดงถึงผลของการโจมตีอย่างไร ทว่า เป็นที่น่าสนใจว่าเราอาจสังเกตเห็นถึงความสัมพันธ์จากการจัดหมวดหมู่ เมื่อเทียบกับข้อมูลเป้าหมายดั้งเดิม และคำตอบเป้าหมายของแบบจำลอง การวิเคราะห์การจัดกลุ่มในลักษณะดังกล่าว หากต่อยอดเพิ่มเติม อาจสามารถบ่งบอกถึงแนวโน้มที่แบบจำลองจะถูกโจมตีได้ (เช่น เลข 3 อาจโจมตีไปเป็นเลข 2 ได้ง่ายเป็นพิเศษ เมื่อดูจากสมาชิกของคลัสเตอร์ที่ 1)

5. สรุป

ผลงานชิ้นนี้มุ่งเน้นถึงการพิจารณาคะแนนการประเมินผล (evaluation) ที่ลดลงของแบบจำลองการเรียนรู้เชิงลึก (deep learning models) ที่ถูกโจมตีประสงค์ร้าย (adversarial attack) ผลการทดลองชี้ให้เห็นว่าอัตราการลดลงของแต่ละเป้าหมาย (target) ไม่เท่ากัน ซึ่งอาจศึกษาเพิ่มเติมได้ว่าเป้าหมายแต่ละเป้าหมายถูกโจมตีได้ด้วยระดับความยากง่ายต่างกัน นอกจากนี้ การวิเคราะห์กลุ่ม (clustering) ยังแสดงให้เห็นถึงแนวโน้มที่กลุ่มของสัญญาณรบกวนจะมีความสัมพันธ์กันในลักษณะที่อาจนำไปสู่การศึกษาเพิ่มเติมได้ว่าชุดของข้อมูลที่มีเป้าหมาย (target) เหมือนกัน จะถูกโจมตีและทำให้แบบจำลองเห็นเป็นเป้าหมายกลุ่มเดียวกัน

เอกสารอ้างอิง

- [1] A. Kurakin, I. J. Goodfellow, and S. Bengio, “Adversarial machine learning at scale,” 2017.
- [2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.
- [3] Z. Kolter and A. Madry, “Adversarial robustness - theory and practice.”
- [4] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” 2014.
- [5] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, “Adversarial examples are not bugs, they are features,” 2019.
- [6] Y. LeCun, C. Cortes, and C. Burges, “Mnist handwritten digit database,” *ATT Labs [Online]*. Available: <http://yann.lecun.com/exdb/mnist>, vol. 2, 2010.