

การจำแนกกลุ่มของสัญญาณโจมตีแบบจำลองการเรียนรู้เชิงลึก

Clustering Analysis of Adversarial Perturbations on Deep Learning Models

ศิระกร ลำไย*, วัชรพัฐ เมตตานันท์†, และ จิตรัท ศัน ฝักเจริญผล‡
ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์
Email: *sirakorn.l@ku.th, †vacharapat@eng.src.ku.ac.th, ‡jtf@ku.ac.th

บทคัดย่อ

บทความวิชาการนี้กล่าวถึงการฝึกสอนแบบจำลองอย่างง่าย เพื่อค้นหารูปแบบของสัญญาณรบกวนที่สามารถโจมตีชุดข้อมูลรับเข้า ให้ได้ผลลัพธ์ของแบบจำลองที่ผิดเพี้ยนไปได้ โดยพิจารณาการโจมตีบนโครงข่ายประสาทเทียมแบบเชื่อมถึงกันทั่ว และโครงข่ายประสาทเทียมแบบสังวัตนาการ โครงข่ายประสาทเทียมทั้งสองแบบถูกฝึกสอนด้วยชุดข้อมูล MNIST ซึ่งมีชุดข้อมูลสำหรับฝึกสอนจำนวน 60,000 จุด หลังจากการฝึกสอนโครงข่ายประสาทเทียม นำข้อมูลทดสอบจำนวน 10,000 จุด มาทำการหาสัญญาณโจมตีความยาวเท่าจำนวนจุดทดสอบ และพยายามทำการเรียนรู้จัดหมวดหมู่ ผลลัพธ์ที่ได้คือ...

คำสำคัญ: ปัญญาประดิษฐ์, จักรกลเรียนรู้, การเรียนรู้เชิงลึก, การโจมตีการเรียนรู้

Abstract

blablabla

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

1. ความสำคัญและที่มา

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

ในการเรียนรู้เชิงตัวแปรเสริม (parameter-based learning) นั้น ตัวแปรเสริม (parameters) ค่าน้ำหนัก (weights) บนแบบจำลองการเรียนรู้เชิงลึก (deep learning models) เป็นตัวกำหนดความฉลาดของแบบจำลอง อาจมีตัวแปรเสริมบางชุดที่ทำให้แบบจำลองมีช่อง

โหวต่อการโจมตีประสงค์ร้าย การโจมตีนั้นอาจเกิดจากการเพิ่มสัญญาณรบกวนซึ่งผ่านการคำนวณ (calculated artefacts) เข้าสู่ข้อมูลรับเข้า (inputs) ซึ่งทำให้ความผิดพลาดของแบบจำลองในการพยากรณ์คำตอบนั้นเปลี่ยนไปอย่างชัดเจน

2. ขบวนการรบกวนและกระบวนการวิธี

2.1. การโจมตีการเรียนรู้

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) ในระดับที่ต่ำกว่าตามนุษย์จะมองเห็น กล่าวคือเมื่อเทียบกับชุดข้อมูลรับเข้าแล้ว ช่วง (range) ของสัญญาณรบกวนนั้นน้อยกว่าช่วงของข้อมูลรับเข้าที่เป็นไปได้มาก การนิยามดังกล่าวจะใช้นิยามเซตของสัญญาณรบกวนที่เป็นไปได้ทั้งหมด (พิจารณาว่ามีค่า η ที่เป็นไปได้หลายค่า และแต่ละค่าโจมตีแบบจำลองได้แตกต่างกันออกไป) ว่า

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

เมื่อนิยามให้ตัวดำเนินการนอร์มอินฟินิตี้ (infinity norm) เป็น

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

และค่า ϵ เป็นค่าคงที่บ่งบอกความเข้มข้นของสัญญาณมากที่สุดที่รับได้โดยมากมิกมีค่าน้อย

2.2. ฟังก์ชันสูญเสีย และการฝึกสอนแบบจำลองด้วยวิธีก้าวเคลื่อนถอยหลัง

พิจารณาการเรียนรู้แบบจำลอง M จะพบว่าการหาตัวแปรเสริม (parameters) θ ที่ดีที่สุดของ M นั้นทำได้ด้วยการนิยามฟังก์ชัน

สูญเสีย (loss function) ℓ_i ของจุดฝึกหัด (training point) i ได้ โดยให้ฟังก์ชันสูญเสียเป็นฟังก์ชันที่เปรียบเทียบเป้าหมาย (target) y_i จากจุดฝึกหัด 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 ในการศึกษาแบบจำลองการเรียนรู้เชิงการจำแนก (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

นอกจากนี้เราอาจนิยามผลรวมของฟังก์ชันสูญเสียทั่วทั้งชุดฝึกสอน

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

เป็นผลรวมของฟังก์ชันสูญเสียบนทุกจุดฝึกหัด เมื่อ N เป็นขนาดของชุดฝึกหัด (training set)

อย่างไรก็ดี แม้สมการ 4 และ 5 จะดูเหมือนพิจารณาค่าสูญเสียที่เปลี่ยนไปเมื่อชุดข้อมูลฝึกหัดเปลี่ยน 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

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

เมื่อ η เป็นค่าอัตราการเรียนรู้ (learning rate) โดยปกติมักมีค่าไม่มาก

หากอธิบายโดยคร่าว ขั้นตอนวิธีการเกรเดียนต์ดลัน พยายามหาค่าตัวแปรเสริม θ_{Opt} โดยการเริ่มจากการสุ่มตัวแปรเสริม θ แล้วค่า

นวนเกรเดียนต์ของฟังก์ชันสูญเสีย และค่อยๆ ปรับค่า θ ตามทิศตรงข้ามกับเกรเดียนต์เรื่อยๆ จนกระทั่งถึงจุดที่ฟังก์ชันสูญเสียมีค่าน้อยที่สุด

2.3. การหาสัญญาณรบกวนด้วยวิธีการก้าวเคลื่อนถอยหลัง

2.4. คำอธิบายต่อการเกิดขึ้นของสัญญาณรบกวน

มีหลายทฤษฎีพยายามอธิบายการเกิดขึ้นของการโจมตีแบบจำลอง ซึ่งอาจยกตัวอย่างทฤษฎีและคำอธิบายได้ดังนี้

2.4.1. การประพัตต์เป็นเส้นตรง. Goodfellow และคณะ พิจารณาลของการโจมตีที่เกิดจาก \tilde{x} อาจพิจารณาได้จากการคุณสมบัติเพื่อหาค่าส่งออกจากชุดน้ำหนัก (weights) ของชั้นแบบจำลองการเรียนรู้เชิงลึก (deep learning layers)

$$w^T \tilde{x} = w^T x + w^T \eta \quad (9)$$

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

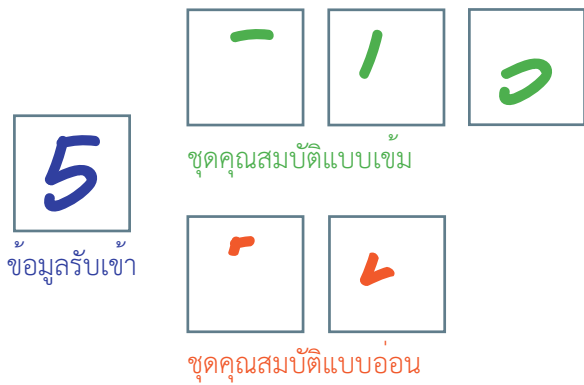
เพื่อพิสูจน์ทฤษฎีดังกล่าว Goodfellow และคณะ พิจารณาผลความน่าจะเป็นของคำตอบที่ออกจากแบบจำลองเมื่อปรับค่า ϵ ดังแสดงในสมการที่ 2 และพบว่าความน่าจะเป็นของข้อมูลส่งออก (output) ของแต่ละชั้นข้อมูล (class) มีความสัมพันธ์เชิงเส้นตรงกับค่า ϵ ที่เพิ่มขึ้นเรื่อยๆ

2.4.2. ทฤษฎีชุดคุณสมบัติแบบอ่อนและแบบเข้ม. Ilyas และคณะ ศึกษาโครงสร้างของแบบจำลองเชิงลึก จนนำมาสู่ข้อสรุปว่า “ช่องโหว่ในการโจมตีแบบจำลองเป็นผลโดยตรงจากความอ่อนไหวของแบบจำลองในการวางหลักการบนชุดคุณสมบัติของข้อมูล” (“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
- ชุดคุณสมบัติแบบอ่อน คือชุดคุณสมบัติที่มนุษย์ไม่สามารถมองเห็น หรือมองเห็นแต่ไม่ได้หยิบมาเป็นตัวบ่งชี้หลักในการตัดสินใจ และวางหลักการในการจำแนก

จะยกตัวอย่างกรณีการจำแนกเลข 5 เราอาจพิจารณาว่าเลข 5 ดังแสดงในรูปที่ 1 ประกอบขึ้นจากขีดหนึ่งขีดแนวขวาง ขีดหนึ่งขีดแนว



รูปที่ 1. ตัวอย่างชุดคุณสมบัติแบบอ่อน และแบบเข้มที่เป็นไปได้ จากเลข 5

ตั้ง และส่วนโค้งคล้ายวงกลม เป็นชุดคุณสมบัติที่มนุษย์สังเกตและเข้าใจโดยทั่วไป รวมถึงเป็นคุณสมบัติที่มนุษย์ใช้ในการสังเกตเห็นเส้นที่เชื่อมต่อกันจนประกอบเป็นเลข 5 อย่างไรก็ตาม แบบจำลองการเรียนรู้ใดๆ อาจเห็นมุมรอยต่อระหว่างขอบ (ซึ่งอาจสังเกตได้ว่าไม่มีเลขตัวใดเลยนอกจาก 1 ถึง 9 ยกเว้น 5 ที่มีมุมและขอบดังแสดง) เป็นตัวตัดสินใจในการเรียนรู้เลข 5 อย่างไรก็ตาม พึงสังเกตว่าแบบจำลองอาจจะแม้กระทั่งเลือกสังเกตเห็นพื้นที่ว่างบริเวณที่แตกต่างกันไป และใช้พื้นที่ว่างเหล่านั้นเพื่อสร้างข้อสรุปหรือตัดสินใจว่าเลขที่มองเห็นเป็นเลขใด (ซึ่งการนำมาซึ่ง “ข้อสรุป” จากที่ว่างนั้น ชัดกับวิสัยปกติของมนุษย์ในการสังเกตและมองเห็นอย่างชัดเจน)

2.5. การทดลองหาสัญญาณรบกวนบนชุดข้อมูล MNIST

2.6. การจำแนกกลุ่มของสัญญาณรบกวน

3. ผลลัพธ์

4. อภิปรายผล

5. สรุป