

การจำแนกกลุ่มของสัญญาณโจมตีแบบจำลองการเรียนรู้เชิงลึก

Clustering Analysis of Deep Learning Adversarial Perturbations

บทคัดย่อ

บทความวิชาการนี้กล่าวถึงการฝึกสอนแบบจำลองอย่างง่าย เพื่อค้นหารูปแบบของสัญญาณรบกวนที่สามารถโจมตีชุดข้อมูลรับเข้าให้ได้ผลลัพธ์ของแบบจำลองที่ผิดเพี้ยนไปได้ 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 อาจมีตัวแปรเสริมบางชุดที่ทำให้แบบจำลองมีช่องโหว่ต่อการโจมตีประสงค์ร้าย การโจมตีนั้นอาจเกิดจากการเพิ่มสัญญาณรบกวน

ซึ่งผ่านการคำนวณ (calculated artefacts) เข้าสู่ข้อมูลรับเข้า (inputs) ซึ่งทำให้ความผิดพลาดของแบบจำลองในการพยากรณ์คำตอบนั้นเปลี่ยนไปอย่างชัดเจน

2. ขบวนการกรรมและกระบวนการวิธี

2.1 การโจมตีการเรียนรู้

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

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

เมื่อเรียก η ว่าสัญญาณรบกวน (perturbations)

ข้อสังเกตที่เกิดขึ้นคือเราจะนิยามชุดสัญญาณรบกวนดังกล่าว ว่ามีความเข้มข้น (intensity) ในระดับที่ต่ำกว่าตามนุษย์จะมองเห็น กล่าวคือเมื่อเทียบกับชุดข้อมูลรับเข้าแล้ว ช่วง (range) ของสัญญาณรบกวนนั้นน้อยกว่าช่วงของข้อมูลรับเข้าที่เป็นไปได้มาก การนิยามดังกล่าวจะใช้การนิยามเซตของสัญญาณรบกวนที่เป็นไปได้ทั้งหมด (พิจารณาว่ามีค่า η ที่เป็นไปได้หลายค่า และแต่ละค่าโจมตีแบบจำลองได้แตกต่างกันออกไป) ว่า

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

เมื่อนิยามให้ตัวดำเนินการนอร์มอนันต์ (infinity norm) เป็น

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

และค่า ϵ เป็นค่าคงที่บ่งบอกความเข้มข้นของสัญญาณมากที่สุดที่รับได้ โดยมากมักมีค่าน้อย

2.2 การหาสัญญาณรบกวนด้วยวิธีการก้าวเคลื่อนถอยหลัง

พิจารณาการเรียนรู้แบบจำลอง M จะพบว่าการหาตัวแปรเสริม (parameters) θ ที่ดีที่สุดของ M นั้นทำได้ด้วยการนิยามฟังก์ชันสูญเสีย

เสีย (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) ค่าดังกล่าวก็จะยิ่งเพิ่มขึ้นเท่านั้น อย่างไรก็ตาม ในการฝึกสอนแบบจำลองการเรียนรู้เชิงการจำแนก (classification) ส่วนมาก มักใช้ฟังก์ชันสูญเสียเป็นฟังก์ชันสูญเสียแบบความวุ่นวายข้ามชั้น (cross entropy loss)

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

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)

2.3 คำอธิบายต่อการเกิดขึ้นของสัญญาณรบกวน

มีหลายทฤษฎีพยายามอธิบายการเกิดขึ้นของการโจมตีแบบจำลอง ซึ่งอาจยกตัวอย่างทฤษฎีและคำอธิบายได้ดังนี้

2.3.1 การประพฤติตัวเป็นเส้นตรง

Goodfellow และคณะ พิจารณาของการโจมตีที่เกิดขึ้นจาก \tilde{x} อาจพิจารณาได้จากการคุณสมบัติการเพื่อหาค่าส่งออกจากชุดน้ำหนัก (weights) ของชั้นแบบจำลองการเรียนรู้เชิงลึก (deep learning layers)

$$w^\top \tilde{x} = w^\top x + w^\top \eta \quad (7)$$

คณะวิจัยสังเกตพฤติกรรมว่าสัญญาณรบกวน η กระตุ้นส่วนของชุดน้ำหนักและฟังก์ชันกระตุ้น (activation function) ในแบบจำลอง

ให้ประพฤติตัวเยี่ยงฟังก์ชันเส้นตรง (linear functions) ซึ่งการแสดงผลพฤติกรรมดั่งเส้นตรง (linearity) ในกรณีขยขอบ (edge case) ของข้อมูลรับเข้านั้นก่อให้เกิดความเป็นไปได้ที่แบบจำลองจะถูกโจมตี

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

2.4 การทดลองหาสัญญาณรบกวนบนชุดข้อมูล MNIST

2.5 การจำแนกกลุ่มของสัญญาณรบกวน

3. ผลลัพธ์

4. อภิปรายผล

5. สรุป