QUANOS- Adversarial Noise Sensitivity Driven Hybrid Quantization of Neural Networks

https://youtu.be/JRrJkgGHWBI

송경주

HSU 한성대학교 HANSUNG UNIVERSITY

CryptoCraft LAB
https://crypto.modoo.at

Contents

CryptoCraft LAB

논문 요약

- 기존 DNN이 Adversarial attack에 취약성을 보임.
- 에너지 절약 및 Adversarial attack을 방어하기 위해 양자화 기술 사용.
- 기존에는 모든 DNN layer을 동일한 비트폭으로 양자화 함.
- 각 Layer별로 Adversarial noise에 대한 민감도 계산
- 각 Layer별 민감도에 따라 에너지 효율적이고 정확하며 adversarial attack에 강력한 하이브리드 양자화 (QUANOS) 제안.
- 벤치 마크 데이터 세트 (CIFAR10, CIFAR100)를 사용하여 QUANOS를 테스트하여 효율성 증명.

→최종적으로 제안하는 기법을 통해 Adversarial attack에 견고할 뿐만 아니라 정확도를 유지하며 모델 크기를 줄였다.

개요

하이브리드 양자화를 통해 Adversarial attack을 방어할 뿐만 아니라 효율적인 에너지 사용과 견고성을 제안.

- 문제 정의

-일반적인 DNN은 Adversarial attack 을 이용해 분류를 잘못하도록 속일 수 있음. (Adversarial attack에 취약성을 보임)

-기존에는 DNN의 모든 layer을 동일한 비트폭으로 양자화 하여 각 계층에 효율적인 양자화를 하지 못함.

→(각 Layer별로 Adversarial noise에 대한 민감도가 다름을 이용한 하이브리드 양자화를 통해 비용절감, 더 견고한 모델 생성)

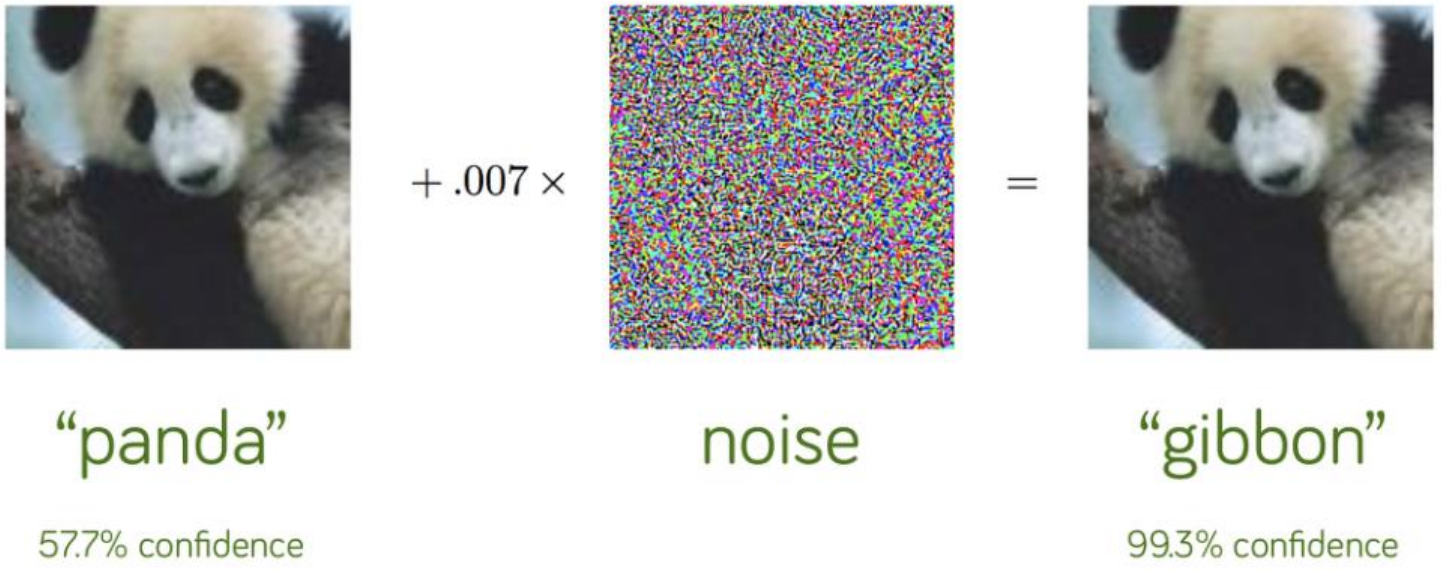
사전 지식

- Background
  1. Adversarial attack (적대적 공격)
  2. 양자화
  3. FGSM
  4. BlackBox (BB) attack
  5. White-Box (WB) attack

Adversarial attack

- 분류 성능이 매우 우수한 DNN을 이용한 Classifier(분류자)들에게 적대적 교란을 적용할 경우 분류 알고리즘들이 쉽게 오분류가 발생할 수 있도록 만드는 것.

  (ex. Noise 섞기, 픽셀에 미세한 변화 주기)



양자화

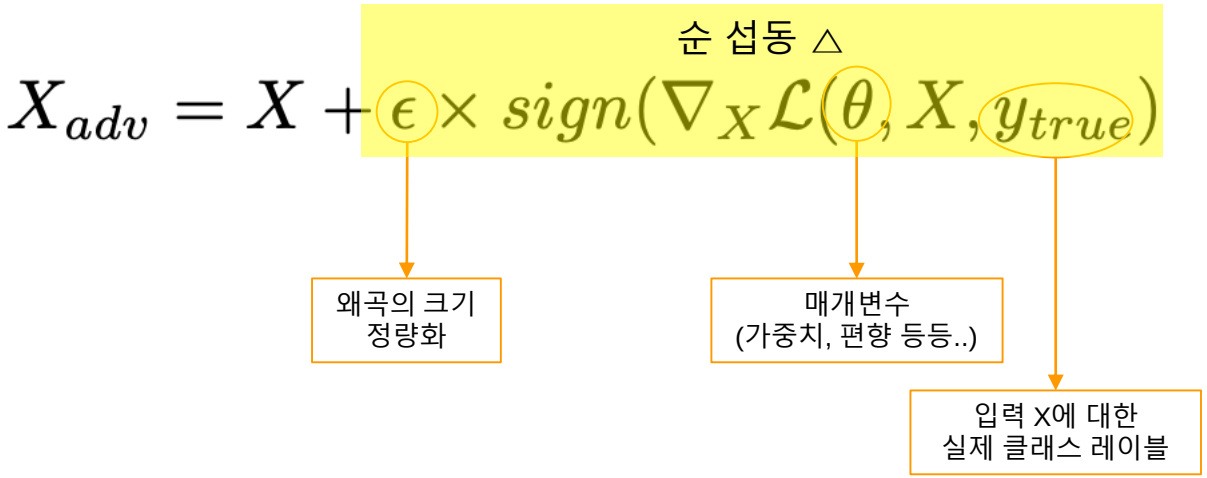
- 정확하고 세밀한 단위로 표현한 입력값을 보다 단순화한 단위의 값으로 변환하는 다양한 기술을 포괄적으로 의미하는 용어

- In DNN

-정보를 잃지 않고 머신 러닝 모델의 크기를 줄일 수 있어 연산량을 줄이면서 전력 효율성을 향상 시키는 방법 중 하나로 사용됨.

FGSM

- Adversarial 공격방법 중 하나. 이를 이용해 Adversarial Example을 만들 수 있음.



BlackBox (BB) attack

- 공격자가 공격 대상 모델의 매개변수에 대한 지식 없이 공격하는 경우.

White-Box (WB) attack

- 공격자가 대상 모델 훈련 정보에 대한 지식을 완전히 가지고 있는 경우, WB의 공격은 매개변수를 사용한다.

- BlackBox(BB) attack보다 강력한 보안 개념이며 WB 공격에 대해 견고성을 보이면 BB 공격에 대해서도 견고성을 보장함.

제안 기법

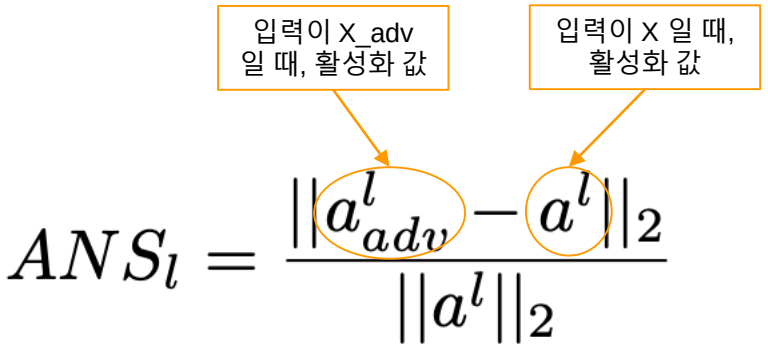
- ANS(Adversarial Noise Sensitivity) : 적대적 소음 민감도를 계산하여 각 계층의 최적의 비트폭을 결정

- QUANOS : ANS에 기반한 Layer 중요도에 따라 각 Layer 별로 양자화 하는 하이브리드 기법.

ANS

- ANS (= 오류율)

- 각 Layer의 adversarial 기여도를 추정하고 최적의 비트 폭을 결정한다.
- 높을수록 적대적 기울기에 더 많이 기여하는 것을 의미하며 많은 변화에 영향을 줌.



QUANOS

- Adversarial 섭동에 대해 Layer 기여도를 기반으로 매개 변수 중요도 할당

→ANS가 낮을수록 높은 정밀도로 양자화 된다.

ANS 값을 이용하여 각 계층 양자화

$$k_l = k_{l_{initial}} - round(ANS_l * k_{l_{initial}})$$

QUANOS

- 비트 압축을 결정하는데 사전 훈련된 모델에 의존하지 않는다. (부분적으로 훈련된 모델을 가져와 QUANOS 분석 수행 → 훈련 에너지 절약)

<QUANOS 반복적 적용>

1. 몇 epoch 동안 DNN 훈련을 함.
2. ANS 계산
3. ANS 기반으로 DNN 양자화
4. 수렴 될 때까지 양자화된 DNN훈련
5. 최적의 정확도, 효율이 달성될 때까지 2-4단계 반복

열심히 해본 결과...
16비트 기준선 ($k_{initial}$ = 16) 에서 시작하면
최적의 DNN이 생성된다는 것을 확인함.

Algorithm 1: QUANOS Procedure

1 Take a randomly initialized DNN (say, 16-bit) and train it for 20-30 epochs;
2 **for** *each layer* l *in DNN* **do**
3 Compute ANS_l using Eqn. 2;
4 $k_l = 16 - round(16 * ANS_l)$ (Eqn. 3);
5 **end**
6 Train the k_l-bit hybrid precision DNN till convergence;

검증

1. ANS value에 따른 adversarial 취약성 평가

- VGG-19 trained on CIFAR10.
- 각 layer 별 ANS 값을 계산하고 ANS 값에 따른 적대적 취약점을 확인하기 위해 계층별로 조금씩 제거하여 절대적 정확도를 확인해 봄.

Q & A