기계학습에 쓰이는 오류정정부호의 부호책의 설계 연구

이선용, 노종선
서울대학교

sonyongyi@snu.ac.kr, jsno@snu.ac.kr

A Study on the Codebook of Error Correcting Output Codes

Lee Seon Yong, No Jong Seon
Seoul National Univ.

요 약

본 논문은 기계학습에서 적대적 공격에 좋은 방어 방법으로 제시되고 있는 Error Correcting Output Codes (ECOC) 에서 사용되는 코드북을 분석한 연구이다. 코드북의 설계에 따라 인공 신경망의 학습결과에 어떤 영향을 미치는지를 파악했다. 이에 따라 향후 ECOC 연구에서 나아갈 방향성을 제시하고 여러 부호책의 장단을 확인했다.

Ⅰ. 서 론

적대적 공격은 최초로 제시된 이후로[1] 학계의 다양한 관심을 끌어왔다. 적대적 공격이란 원본 데이터에 인간이 인식할 수 없는 약간의 잡음을 추가하는 것으로 인공 신경망이 오분류를 일으키도록 하는 공격 방법이다. 이는 기계학습의 매우 치명적인 문제로 알려져 다양한 공격방법과 방어방법이 제시되고 있다. 익히 알려져 있는 공격방법으로는 표준 공격방법으로 인정받고 있는 Projected Gradient Descent (PGD) 공격[2], Fast Gradient Sign Method (FGSM)[1], 매우 강력한 공격 방법인 Carlini-Wagner 공격[3] 등이 있다. 이에 대처하는 방어 방법으로는 적대적 학습[2], 앙상블 인공신경망, ECOC[4] 등이 제시되고 있다.

ECOC 는 적대적 학습 다음으로 각광받고 있는 방어 방법이다. ECOC 는 인공신경망의 활성 신경층을 기존의 Sigmoid 함수 대신 tanh 함수를 이용하고, Softmax 함수를 Correlation 값을 뽑는 방식 (1)으로 대체하였다.

$$\mathrm{p}_\sigma(k) = \frac{\max((\sigma(z)*C_k,0))}{\sum_{i=1}^{M}(\max(\sigma(z)*C_i,0))} \quad (1)$$

이를 통해 인공 신경망의 출력값과 부호책의 부호 사이의 correlation 을 계산하여 가장 비슷한 값을 계산하는 방식으로 동작한다. 최초의 ECOC 에서는 부호책을 하다마드 부호를 이용했다. 하다마드 부호는 최소해밍거리를 최대한으로 확보할 수 있는 (2^n-1) 부호이기 때문에 사용되었고 이를 기준으로 다양한 코드북 설계가 연구되고 있다.

본 논문에서는 다양한 부호책들을 이용 적대적 공격에 대한 결과와 몇가지 특성들을 분석했다.

Ⅱ. 본론

본논문에서는 ECOC 에서 사용되는 부호책들을 몇 종류 변경해가며 MNIST, CIFAR-10 데이터에 대해 실험을 진행했다.

부호책은 두 가지 방법으로 만들었다. 첫번째 부호책은 하다마드 부호책을 기반으로 하여 channel capacity 를 변화하지 않는 선에서 random bit flip 을 적용한 부호책을 사용하였다.

두번째 부호책은 random bit flip 이 진행된 matrix 를 바탕으로 각 class 별 accuracy 를 측정한 뒤, class accuracy 가 높은 class 들에 average hamming distance 가 낮은 부호를 부여한 부호책을 사용하였다.

Table 1 MNIST Result

Model	*Clean*	*FGSM*	*PGD*
Standards	99.18	15.56	0.08
Madry[2]	99.14	96.1	93.68
ECOC[4]	99.47	96.8	95.8
Mymethod1	99.45	95.8	93.59
Mymethod2	99.38	96.4	95.22

Table 2 CIFAR-10 Result

Model	*Clean*	*FGSM*	*PGD*
Standards	86.44	10.02	0.01
Madry[2]	87.20	56.10	45.80
ECOC[4]	87.64	60.80	55.40
Mymethod1	88.81	61.59	56.60
Mymethod2	87.10	62.60	56.73

Ⅲ. 결론

본 논문에서는 기존의 ECOC 에서 사용하던 부호책에 random bitflipping 과 class accuracy 를 바탕으로 부호책 설계에 변형을 주어 그에 따른 정확도를 cifar-10 과 mnist 데이터에 대해 실험을 진행했다. 결과 class accuracy 를 바탕으로 부호책을 설계했을 때 좋은 결과가 나와 향후 연구에서 부호책 설계에 연구할 방향이 다양할 것이다.

.

ACKNOWLEDGMENT

이 논문은 2021 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2021-0-00400, 저사양 디바이스 대상 고효율 PQC 안전성 및 성능 검증 기술 개발)

참 고 문 헌

[1] Ian J. Goodfellow, Jonathon shlens, Christian Szegedy " Explaining and Harnessing Adversarial Examples," 2015. [Online]. https://arxiv.org/abs/1412.6572

[2] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, Adrian Vladu, Towards Deep Learning Models Resistant to Adversarial Attacks," 2018. *The International Conference on Learning Representations*

[3] Nicholas Carlini, David Wagner. "Towards Evaluating the Robustness of Neural Networks," 2017. IEEE Symposium on Security and Privacy

[4] Gunjan Verma, Ananthram Swami, “Error Correcting Output Codes Improve Probability Estimation and Adversarial Robustness of Deep Neural Networks,” 2019. 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.