오류정정부호를 이용한 인공신경망에 대한 적대적 공격에 대한 적대적 공정성에 대한 연구

이선용, 노종선
서울대학교 전기정보공학부 뉴미디어통신공동연구소

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

Error Correcting Output Code Network and Adversarial Robust Fairness

Lee Seon-Yong, No Jong-Seon
Dept. of ECE, Seoul National University

요 약

본 논문은 오류정정부호가 인공신경망에서 적대적 공정성에 미치는 영향에 대하여 분석하였다. 오류정정부호의 해밍거리 조절이 적대적 공정성에 영향을 줌을 확인하고 그에따른 효과를 분석했다. 결과 해밍거리 조절을 통해 적대적 공정성에 상대적 이득을 취할 수 있음을 확인했다.

.

Ⅰ. 서 론

본 논문에서는 오류정정부호, 인공신경망, 적대적 공격, 적대적 공정성에 관한 내용을 다룬다. 인공 신경망들을 여러 개 사용하여 신경망의 성능을 높이는 방법에는 여러가지가 있지만 그중 매우 유명한 것이 ensemble 방법이다. 여러 개의 인공신경망을 평행, 혹은 순차적으로 배치함으로 단일 네트워크보다 성능을 높이거나 적대적 공격에 대한 강건함을 증가시킬 수 있다.

오류정정부호 인공신경망은 인공 신경망에 오류정정부호를 적용한 인공 신경망으로 상황에 따라 이전에 언급한 ensemble 방법을 적용한 방법과, 적용하지 않은 두 가지 방법 모두에 적용할 수 있다. 오류 정정부호를 적용하는 방법은 ditterich [1] 의 논문에서 제시되고 발전되었다. 다중분류기를 여러 개의 이진분류기의 연합을 통해 구성하는 방식이다. 이에 하다마드 부호를 적용하여 적대적 공격에 강건한 인공신경망을 구성하는 방법이 제시되었으며 [2], 하다마드 부호 이외에도 다른 여러 부호를 이용한 네트워크를 만드는 방법들이 제시되었다.

적대적 공정성이란 인공신경망의 분류 도중 클래스마다 적대적 공격을 가했을 때와 가하지 않았을 때의 정확도의 차이가 달라지는 것을 의미한다. 적대적 공정성을 높임으로 적대적 공격에 대한 저항성을 높이는 연구가 제시되고 있다.

본 논문에서는 해밍거리 조절이 적대적 공정성에 주는 영향에 대해 분석했다.

Ⅱ. 본론

하다마드 부호는 모든 행,열 간에 해밍거리를 최대로 만들 수 있는 행렬이다. 행렬의 크기가 $2^k \times 2^k$ 일 때 해밍거리는 2^{k-1} 을 가진다. 이 하다마드 부호에 대해 무작위 비트 뒤집기를 행해, 최소해밍거리가 8 인 $k = 4$ 의 하다마드 행렬을, 해밍거리가 7~11 사이의 값을 가지도록 행렬을 변환해준다.

총 세개의 행렬을 바탕으로 해밍거리와, 공정성을 파악했다. 그 행렬은 각각 다음과 같다.

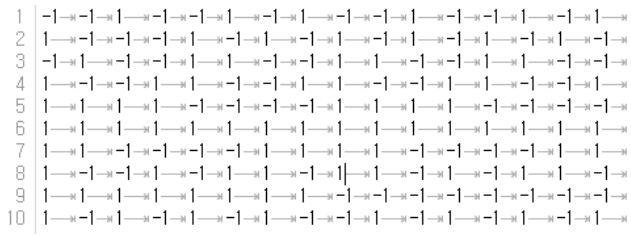

Figure 1. 행렬 1

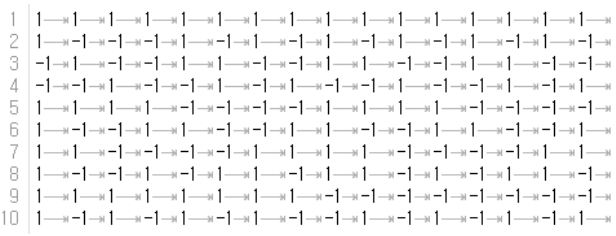

Figure 2 행렬 2

```
1  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
2  1 -1 -1 -1 1 -1 -1 -1 1 -1 -1 -1 1 -1 -1 -1 1 -1 -1 -1
3  -1 1 -1 -1 1 1 -1 -1 1 1 -1 -1 1 1 -1 -1 1 -1 -1 -1
4  1 -1 -1 1 1 -1 -1 1 1 -1 -1 1 1 -1 -1 1 -1 -1 1
5  1 1 1 1 -1 -1 -1 -1 1 1 1 1 -1 -1 -1 -1
6  1 1 1 1 1 1 1 1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
7  1 1 -1 -1 -1 -1 1 1 1 1 -1 -1 -1 -1 1 1 1
8  1 -1 -1 1 1 -1 1 1 -1 1 1 1 -1 1 -1 -1 1 1 -1
9  -1 -1 1 1 -1 -1 1 1 -1 1 -1 -1 1 1 -1 -1 1 1 -1 1
10  1 -1 1 1 -1 1 -1 1 -1 -1 1 1 -1 1 -1 -1 1 1 -1 1
```

Figure 3 행렬 3

그리고 그에 따른 혼동 행렬, 클래스별 정확도, 적대적 공격 상황일 때 와 아닐때의 정확도 차이를 계산했다.

```
[[919   7  17   0   6   0   2   6  30  13]
 [  4 958   1   1   0   3   1   3   7  22]
 [ 48   3 847  31  18  18  12  11   5   7]
 [ 23   4  49 731  27 105  16  21  12  12]
 [ 17   0  40  35 832  22   9  38   5   2]
 [  4   3  26  97  17 816   6  26   3   2]
 [  8   4  30  42  13  11 877   8   3   4]
 [  8   1   6  15  18  22   1 923   1   5]
 [ 27  11   5   4   2   0   0   2 936  13]
 [ 13  34   2   5   0   2   1   4  14 925]]
[0.919 0.958 0.847 0.731 0.832 0.816 0.877 0.923 0.936 0.925]
[[18  3 18  3  6  3  1  3 29  9]
 [17 24  1  1  3  4  1  2 17 39]
 [11  4 47  6 11  4  4  5  0  3]
 [ 7  1  4 26 15 23  6  6  5  3]
 [11  3 28 16 26  2  4 15  3  0]
 [ 2  3 21 25  9 27  0  9  1  2]
 [ 3  3 13 15 15  9 40  3  1  2]
 [11  3 10  7 17 24  0 17  0  2]
 [27  4  5  0  3  4  0  0 66  2]
 [ 5 25  0  2  2  2  1  1 12 44]]
[0.19354839 0.22018349 0.49473684 0.27083333 0.24074074 0.27272727
 0.38461538 0.18681319 0.59459459 0.46808511]
[0.72545161 0.73781651 0.35226316 0.46016667 0.59125926 0.54327273
 0.49238462 0.73618681 0.34140541 0.45691489]
```

Figure 4 행렬 1 의 혼동 행렬, 클래스별 정확도, 적대적 공정성

```
[[918  10  10   2   7   2   1   6  30  14]
 [  3 952   2   1   1   2   0   0   9  30]
 [ 51   2 843  24  26  21   9  12   6   6]
 [ 24   1  40 745  26 106  20  19  10   9]
 [ 21   1  34  22 854  19  14  24  10   1]
 [ 11   2  22  88  22 820   2  28   3   2]
 [ 10   4  29  31  13  19 876  10   4   4]
 [ 12   0  11   6  21  31   2 910   3   4]
 [ 32   4   6   6   2   0   1   0 939  10]
 [ 21  37   4  10   1   2   1   3  20 901]]
[ 8.2  4.8 15.7 25.5 14.6 18.  12.4  9.   6.1  9.9]
[[33  2 12  7 15  3  1  1 13  6]
 [ 5 63  0  3  1  0  3  5 12 17]
 [ 9  1 50 12  4  2  9  5  2  1]
 [ 3  5 10 36  4 22  3  5  3  5]
 [ 2  0 10 25 37  2  7 17  6  2]
 [ 0  2  7 26  3 51  1  9  0  0]
 [ 1  1  6 29  7  3 46  5  1  5]
 [ 4  2  7 16 11  7  1 39  2  2]
 [44  5  2  8  3  1  0  1 44  3]
 [ 9 14  1  3  1  2  2  5  6 51]]
[64.51612903 42.20183486 47.36842105 62.5         65.74074074 48.48484848
 55.76923077 57.14285714 60.36036036 45.74468085]
[0.56316129 0.37401835 0.31668421 0.37         0.51140741 0.30484848
 0.43369231 0.48142857 0.5426036  0.35844681]
```

Figure 5 행렬 2 의 혼동 행렬, 클래스별 정확도, 적대적 공정성

```
[[940   4   8   3   7   1   0   9  19   9]
 [  8 949   2   3   3   1   1   1   4  28]
 [ 50   0 851  34  16  22   4  16   1   6]
 [ 30   4  38 773  39  70  15  20   6   5]
 [ 19   1  56  24 845  13   5  33   3   1]
 [ 22   1  23 126  31 774   1  19   0   3]
 [ 16   3  52  42  22  18 838   3   2   4]
 [ 13   1   9  17  18  16   1 923   0   2]
 [ 47   5  10   6   0   2   0   2 915  13]
 [ 21  39   4   4   1   1   1   1  17 911]]
[0.94  0.949 0.851 0.773 0.845 0.774 0.838 0.923 0.915 0.911]
[[37  5 15  2  9  6  1  0 10  8]
 [16 33  2  4  0  4  1  2 16 31]
 [14  1 41  9 10  3  9  4  3  1]
 [ 1  4 10 32 11 23  6  5  2  2]
 [ 4  1 29  7 25 16  9 13  3  1]
 [ 2  2 16 27  5 32  4  8  3  0]
 [ 3  5 30  9 11  9 25  6  0  6]
 [12  2 16 15  4 23  3 10  3  3]
 [65  4  6  1  1  3  2  2 15 12]
 [ 4  6  0  0  0  2  0  1  3 78]]
[0.39784946 0.30275229 0.43157895 0.33333333 0.23148148 0.32323232
 0.24038462 0.10989011 0.13513514 0.82978723]
[0.54215054 0.64624771 0.41942105 0.43966667 0.61351852 0.45076768
 0.59761538 0.81310989 0.77986486 0.08121277]
```

Figure 6 행렬 3 의 혼동 행렬, 클래스 별 정확도, 적대적 공정성

행렬 1 의 경우 3,4 번 클래스에 대해 각가 평균 해밍거리가 증가하거나 감소함에 따라 공정성이 증가 혹은 감소함을 볼 수 있고, 다른 행렬 2,3 의 경우에도 일치한다.

Ⅲ. 결론

해밍거리에 무작위 변화를 주어서 적대적 공정성에 대한 실험 결과를 확인하였다 결과 특정 라벨간의 해밍거리가 늘어났을 경우, 두 라벨간의 적대적 공정성이 상승하는 것을 확인할 수 있었다. 또한 평균 해밍거리가 증가함에 따라 그 클래스의 적대적 공격 정확성과 일반 정확성의 차이가 줄어듦을 알 수 있다.

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] T.G. Dietterich and G. Bakiri,"Solving Multiclass Learning Problems via Error-Correcting Output Codes", 1995. Jan. Journal of Artifical Intelligence Research.

[2] Gunjan Verma and 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.

[3] Y. Song, Q. Kang, and W.P. Tay, "Error-Correcting Output Codes with Ensemble Diversity for Robust Learning in Neural Networks", Association for the Advancement of Artificial Inteligence 2021.