이진 인공 신경망 검증하기

고현수

hsgo@ropas.snu.ac.kr

2019. 04. 26 SnT

나쁜 공격, 적대적 교란

Adversarial Perturbations

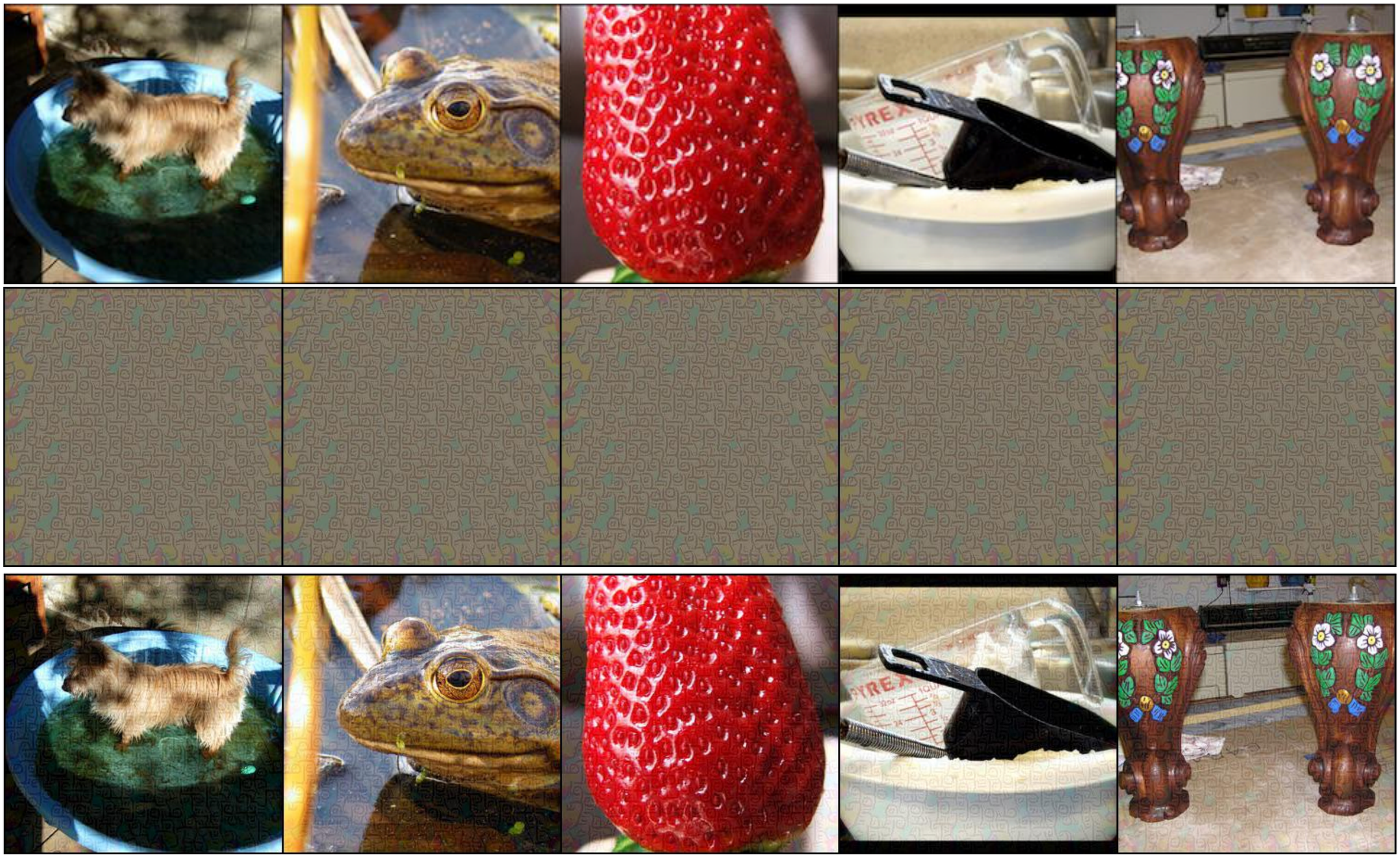
- 신경망이 올바르게 분류하는 **입력**을
  잘못 분류하게 만드는 아주 작은 **변화 혹은 교란**

- 많은 경우, 사람은 변화를 인지하지 하지조차 못함

- 어느 정도 **전이 가능**(Transferable)

예시

[O. Poursaeed *et al*., 2018]



Target: Jigsaw Puzzle, Top-1 target accuracy: 89.3%

전이 가능성

Transferability

- 한 신경망을 교란하는 입력은 다른 신경망도 교란시킬 가능성
- 전혀 다른 구조의 기계 학습 모델로도 전이 가능
  예) DNN -> Decision Tree, kNN…
- 학습 데이터가 달라도 가능

이진 신경망

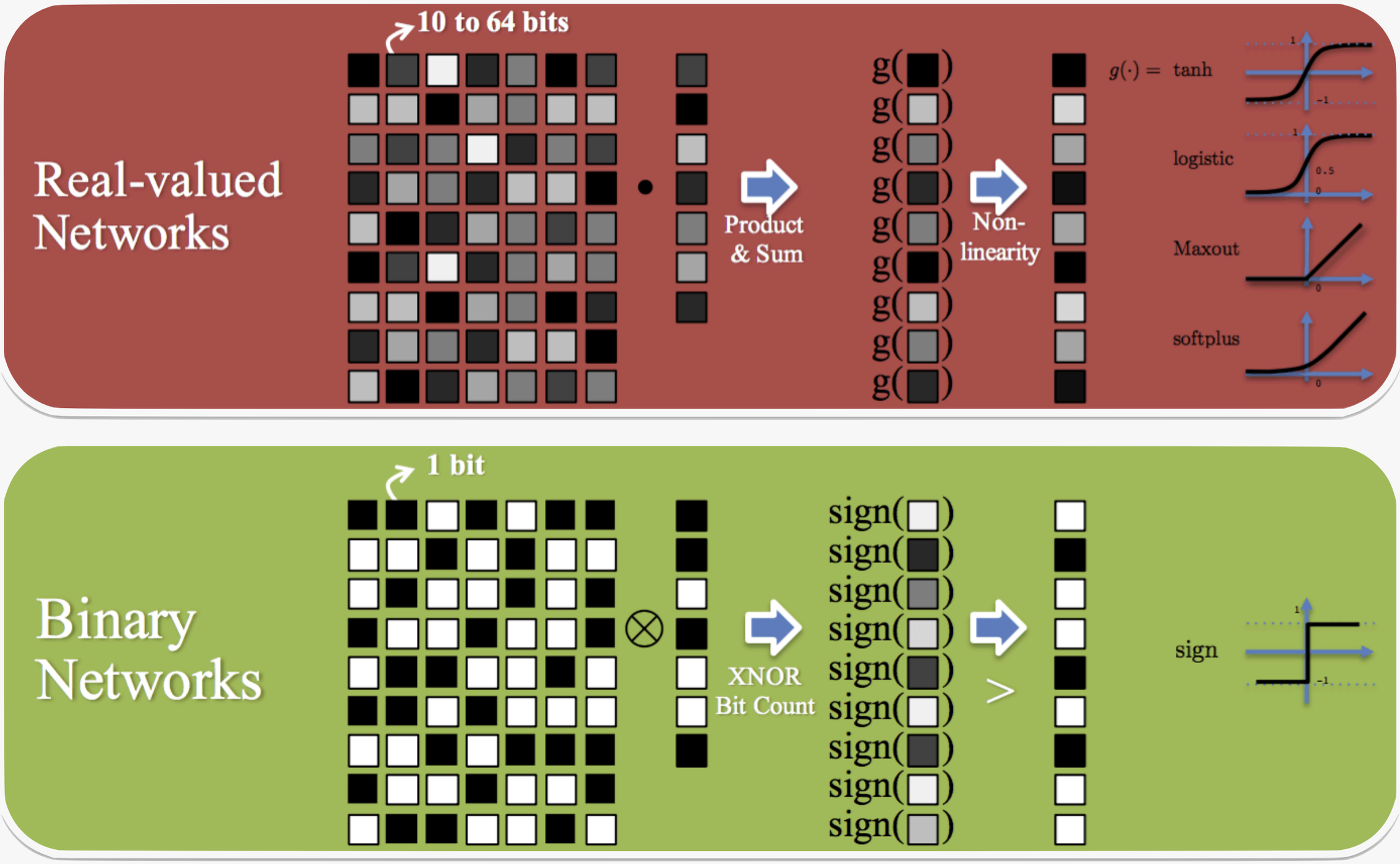
Binary Neural Network

- 실행 중 가중치와 활성 함수의 결과가 모두 **+1** 혹은 **-1**
- 장점
  - 가중치 표현에 32-bit 실수 대신 1-bit 만 사용
  - 값싼 1-bit 연산

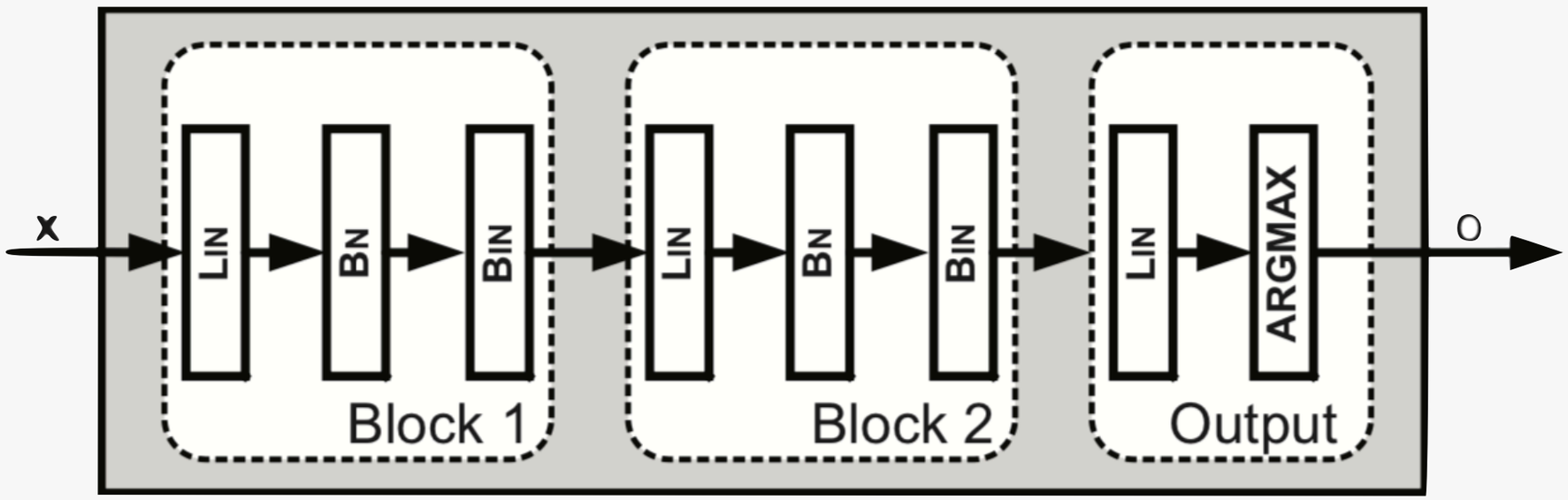
이진 신경망

Binary Neural Network

picture from Deep Learning Resources



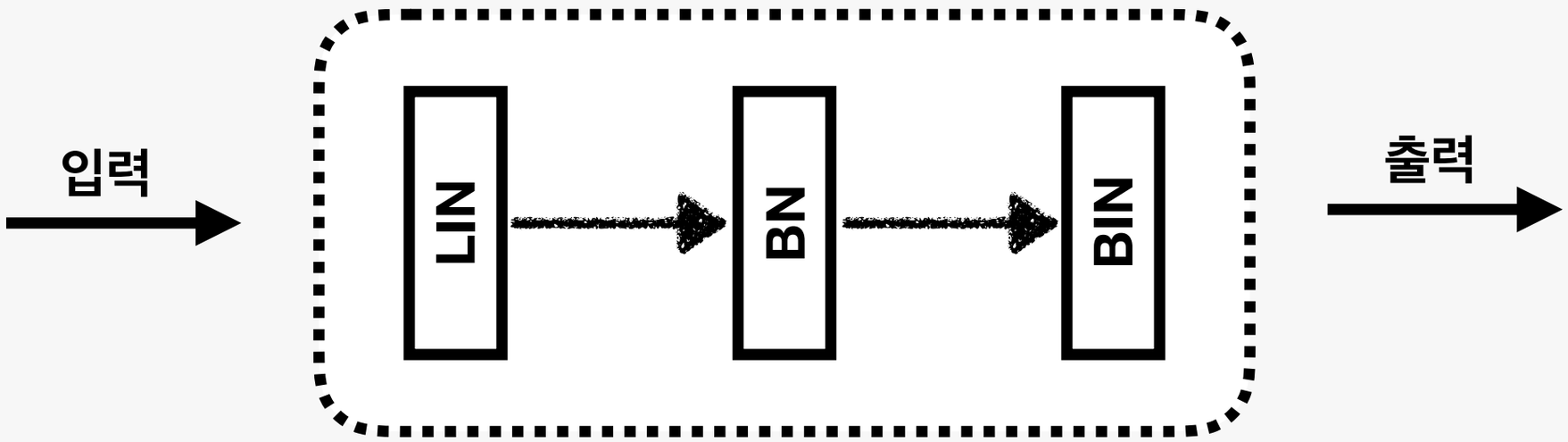
이진 신경망의 구조



[Nina Narodytska *et al*., 2018]

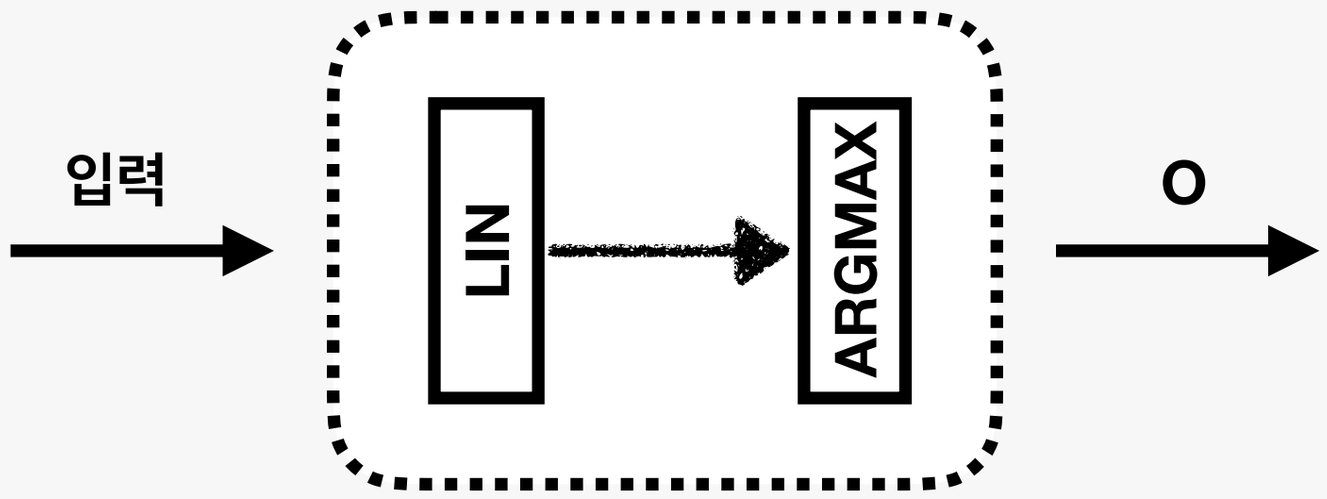
- 여러 개의 내부 블록들
- 출력 층

이진 신경망의 내부 블록



- Linear Transformation
- Batch Normalization
- Binarization
- 입력과 출력은 이진(모두 +1 혹은 -1)

이진 신경망의 출력층



- Linear Transformation
- ARGMAX

이진 신경망 검증

- 이진 신경망을 CNF 로 변환
  - $\left(\bigwedge_{k=1}^{d-1} \mathbf{BINBLK}_k(\mathbf{x}_k, \mathbf{x}_{k+1})\right) \wedge \mathbf{BINO}(\mathbf{x}_d, o)\,.$
- 검증하려는 성질을 CNF로 변환하여 SAT Solver 이용
  - 이 경우,

$$\begin{aligned}\mathbf{BNN}_{Ad}(\mathbf{x}+\tau, o, \ell_{\mathbf{x}}) = \mathbf{CNF}(\|\tau\|_\infty \leq \epsilon) \wedge \\ \bigwedge_{i=1}^{n} \mathbf{CNF}((\mathbf{x}_i + \tau_i) \in [\mathrm{LB}, \mathrm{UB}]) \wedge \\ \mathbf{BNN}(\mathbf{x}+\tau, o) \wedge \mathbf{CNF}(o \neq l_X).\end{aligned}$$

[Nina Narodytska *et al.*, 2018]

검증 대상 모델

MNIST DATASET

picture from Adam Geitgey

검증 대상 모델

- 이진 신경망
- 4개의 내부 블록
    - 첫 블록은 200개, 나머지는 100개의 뉴런들
- 출력층
- 약 95%의 정확도

실험 결과

- 굉장히 복잡한 SAT Encoding
  - 평균 1.4m 개의 변수
  - 평균 5m 구절
- ε=1 일 때, 200 개의 실험 중 180개 해결 (시간제한 300초)

원본 이미지*(위)*와 *SAT*로 찾은 교란된 이미지*(아래)*.

[Nina Narodytska *et al*., 2018]

결론

- 현재 방법은 충분히 큰 모델에는 적용하기 어려움(Scalability)
- **정적분석** 활용
    - 같은 모델을 프로그램으로 나타내기