DPLL(T) 기반 ReLU 신경망 검증을 위한 교육용 오픈소스 소프트웨어 개발

학부생 김송헌*, 노정균*, 이찬희*, *대학원생* 최주언*,
교수 최광훈*, 김재권**, 임효상**

Development of Educational Open-Source Software for DPLL(T)-Based ReLU Neural Network Verification

Undergrad. Songheon Kim*, JeongGyun Noh*, Chanhui Lee* , *Grad.* Jueon Choi*, *Prof.* Kwanghoon Choi*, Jaekwon Kim**, Hyo-Sang Lim**

요 약

본 논문은 DPLL(T) 기반 ReLU 신경망 검증을 위한 교육용 오픈소스 소프트웨어를 제안한다. 본 소프트웨어는 신경망과 입력·출력 조건을 논리식으로 인코딩하고, 명세를 위반하는 반례의 존재 여부를 판정한다. XOR 문제를 학습한 ReLU 신경망을 대상으로 강건성 검증을 수행하여, 강건성이 유지되는 임계 구간과 결정 경계 부근의 구체적 반례를 도출함으로써 소프트웨어의 유효성을 확인하였다.

I. 서론

최근 신경망은 다양한 문제에서 높은 성능을 보이며 널리 활용되고 있으나, 블랙박스 구조로 인해 안전이 중요한 시스템에서 형식적 보장이 어렵다는 한계가 있다. 그 예시로 Eykholt et al.[1]은 정지 표지판에 단순한 스티커를 부착하는 것만으로도 교통 표지판 인식 시스템이 이를 속도 제한 표지판으로 오분류함을 보였다. 이는 미세한 물리적 변형이 모델의 예측을 크게 변화시킬 수 있음을 보여주며, 실제 환경에서 신경망 강건성 검증의 필요성을 시사한다.

Katz et al.[2]은 무인 항공기 충돌 회피 시스템(ACAS Xu)의 ReLU 신경망을 대상으로, 신경망의 동작을 논리식으로 인코딩하고 그 만족 가능성을 판정하는 SMT 기반 검증 도구 Reluplex를 제안하여, 신경망 검증 문제를 체계적으로 풀 수 있음을 최초로 보였다.

이후 신경망 검증은 빠르게 발전하여 2020년부터 매년 국제 경진대회(VNN-COMP)[3]가 개최되고 있다. 5년 연속 1위인 α,β-CROWN[4]은 선형 완화 기반 교란 분석(LiRPA)과 분기한정법을 결합한 경계 전파 방식으로, GPU 가속을 통해 대규모 신경망을 효율적으로 검증하나, 지원하는 명세 형식이 제한되어 유연성이 제한된다. 반면 2위인 NeuralSAT[5]는 DPLL(T) 프레임워크를 신경망 검증에 적용한 접근법이다. DPLL(T)는 명제 논리 수준의 탐색(DPLL)과 산술 제약을 처리하는 이론 솔버를 결합한 구조로, 논리식 기반 명세를 통해 다양한 검증 속성을 유연하게 표현할 수 있다. DPLL(T)의 기반인 DPLL 알고리즘은 1962년에 제안된[6] 이래 60년 이상의 연구 역사를 가지며, Microsoft Research의 Z3[7] 등 범용 SMT 솔버의 핵심 엔진으로 활용되고 있어, 확립된 이론적 기반과 향후 발전 가능성을 갖춘 방법론이다.

그러나 신경망 검증은 비교적 최근에 발전한 분야로, 복합적인 구조로 인해 입문자가 전체 과정을 이해하기 어렵고, 이를 통합적으로 학습할 수 있는 자료와 실습 환경도 제한적이다.

본 연구는 과학기술정보통신부의 재원으로 한국연구재단의 지원을 받아 수행되었으며(RS-2025-24523420), 한국인터넷진흥원(KISA) 정보보안 특성화대학 지원사업의 지원을 받아 수행되었다. 또한 본 연구는 과학기술정보통신부의 재원으로 정보통신기획평가원(IITP)의 정보보호핵심원천기술개발사업(RS-2025-25394739) 및 인공지능융합혁신인재양성사업(IITP-RS-2023-00256629)의 지원을 받아 수행되었다.

교신저자 : 최광훈

* 전남대학교(Chonnam National University)

** 연세대학교(Yonsei University)

따라서 본 논문에서는 Albarghouthi[8]가 제시한 신경망 검증 파이프라인, 즉 신경망과 검증 명세를 논리식으로 인코딩하고 DPLL(T) 프레임워크에 따라 만족 가능성을 판정하는 과정을 구현한 교육용 오픈소스 소프트웨어를 제안한다. 본 시스템은 논리식 인코딩부터 DPLL(T) 탐색, 반례 도출까지의 전 과정을 추적할 수 있어, 신경망 검증 이론의 학습을 지원하는 실습 도구로 활용할 수 있다.

II. 제안 시스템 및 실험 결과

2.1 제안 방법

본 시스템의 입력은 신경망의 동작과 검증 명세를 형식적으로 기술한 논리식이다. 구체적으로, 신경망 f의 입출력 관계를 선형 실수 산술(LRA) 공식으로 인코딩한 신경망 모델(F_{NN})과, 입력 조건(Precondition) 및 출력 조건(Postcondition)으로 구성된 검증 명세가 주어진다.

$$(Pre \wedge F_{NN}) \rightarrow Post \qquad (1)$$

검증은 명세가 항상 참임을 직접 증명하는 대신, 해당 명세의 부정을 취한 반례 탐색 질의의 만족 가능성 여부를 확인한다.

$$Pre \wedge F_{NN} \wedge \sim Post \qquad (2)$$

이러한 반례 탐색 질의는 하나의 논리식으로 인코딩된다. 이때 신경망의 각 레이어 연산은 선형 부등식과 ReLU 제약의 조합으로 표현되며, 입력 조건과 출력 조건의 부정을 결합하여 통합된 형태의 논리식으로 구성된다. 또한, 생성된 논리식은 Boolean 수준의 효율적인 탐색을 위해 절들의 논리곱 형태(CNF)로 변환되어 표현된다.

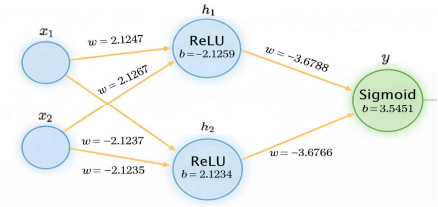
구성된 논리식의 만족 가능성은 DPLL(T) 프레임워크를 통해 판정된다. DPLL(T)는 Boolean 수준의 논리 탐색과 Theory 수준의 제약 검증을 결합하여 복합적인 제약 문제를 해결하는 방법이다. Boolean 수준에서는 각 제약을 하나의 참/거짓 값을 갖는 변수로 단순화한 뒤 DPLL 기반 탐색을 수행하여 가능한 제약 조합을 결정한다.

이후 Theory 수준에서는 선택된 제약 조합을 Reluplex에 전달하여 실제 산술 및 ReLU 의미론에 따라 만족 가능성을 평가한다. 이 과정에서 ReLU 제약 위반이 발생할 경우 값 수정(local repair)을 시도하며, 반복적으로 해결되지 않을 경우 분기 탐색을 수행한다.

최종적으로 시스템은 질의가 SAT인 경우, 입력 조건을 만족하면서 출력 조건을 위반하는 반례가 존재함을 의미하며, UNSAT인 경우 주어진 입력 범위 내에서 검증 명세가 항상 성립함을 의미한다.

2.2 실험 설계

본 절에서는 2.1에서 정의한 검증 방법을 그림 1에 나타낸 XOR 문제를 학습한 ReLU 신경망에 적용한다.



* 그림 1. XOR 신경망의 구성

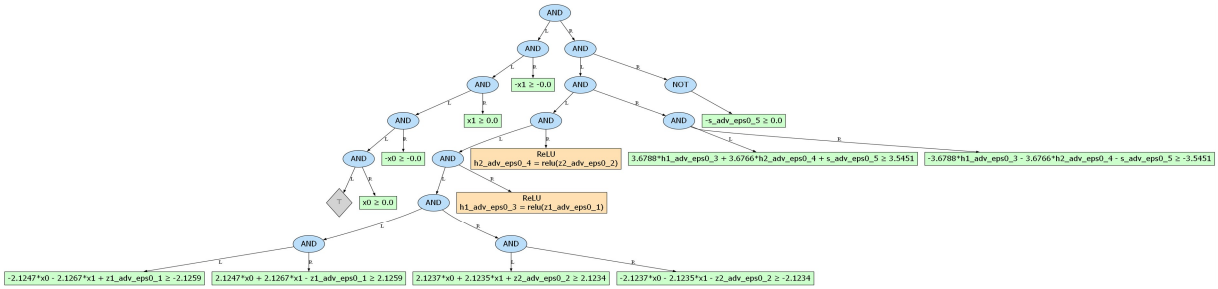
해당 네트워크는 Hidden Layer에서 Affine 변환과 ReLU를 적용하고, Output Layer에서는 최종 Logit을 계산한다. 본 실험에서는 Sigmoid의 단조성에 근거하여 Sigmoid를 직접 인코딩하지 않고 Logit의 부호로 클래스를 판정한다.

본 실험에서는 XOR의 네 가지 중심 입력 (0, 0), (0, 1), (1, 0), (1, 1) 각각에 대해 중심점 주변의 섭동 범위를 설정하고, ϵ값을 변화시키며 반례 탐색을 수행하였다.

강건성 검증을 수행하기 위해서는 특정 중심점 x_c를 기준으로 교란 반경 ϵ을 전제조건 Pre로 정의하고, 해당 범위 내의 모든 입력이 중심 입력과 동일한 분류 결과를 유지하는지 확인한다. 이를 위해 중심점별로 수식 (2) 형태의 반례 탐색 질의를 구성하고, DPLL(T) 솔버를 통해 만족 가능성을 판정하였다. 또한 ϵ을 점진적으로 증가시키며 검증을 반복 수행함으로써, 강건성이 유지되는 임계 범위와 오작동이 시작되는 전이 구간을 분석하였다.

예를 들어, 중심 입력 (0, 0)인 Case에 대해 $Pre: x_1, x_2 \in [0-\epsilon, 0+\epsilon]$, $Post: s < 0$와 같은 명세를 정의하여 이 명세를 위반하는 반례의 존재 여부를 검사한다. 그림 2는 해당 명세

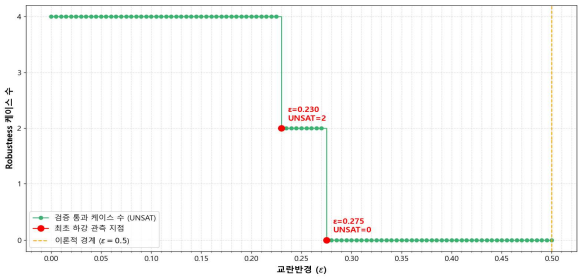
와 신경망 제약 F_{NN}을 결합하여 논리식으로 인코딩한 결과를 트리 구조로 나타낸 것이다.



* 그림 2. XOR 신경망 (0, 0) 입력에 대한 반례 탐색 질의의 논리식 트리

2.3 실험 결과

교란 반경 ε의 증가에 따른 강건성 변화 양상을 분석하기 위해, 각 ε값에서 반례가 발생한 중심 입력 Case의 수를 측정하였다.



* 그림 3. ε 변화에 따른 XOR 신경망의 강건성 검증 결과

그림 3은 ε값의 변화에 따라 반례가 존재한 중심 Case 수의 변화를 나타낸다. x축은 교란 반경 ε, y축은 Robustness 케이스, 즉 각 ε에서 반례가 존재하지 않아 UNSAT이 발생한 케이스의 수이다.

$\varepsilon \leq 0.22$ 구간에서는 네 개의 입력 Case 모두에서 UNSAT이 나타나, 해당 범위에서는 신경망이 강건하게 동작함을 확인하였다. 그러나 $\varepsilon \approx 0.23$ 부근부터는 두 개의 Case에서 SAT이 발생하여 일부 영역에서 강건성이 보장되지 않는다. $\varepsilon \approx 0.275$ 이상에서는 네 개의 모든 Case에서 SAT이 발생하여, 이 범위부터는 더 이상 강건성이 보장되지 않음을 확인하였다.

임계 구간에서 추출한 반례 $x_1 = 0.27085017$, $x_2 = 0.27500000$을 실제 신경망에 대입한 결과, logit 값이 0.000001로 나타나 기대 클래스와 다른 출력을 생성하였다. 이는 결정 경계 부근에서 매우 작은 입력 변화만으로도 출력 클래스가 반전될 수 있음을 보여주며, 그래프상의 SAT 전이 구간이 실제 오작동 경계와 일치함을 확인하였다.

III. 결론

본 논문에서는 Albarghouthi[8]가 제시한 신경망 검증 파이프라인을 구현한 교육용 오픈소스 소프트웨어를 제안하고, XOR 신경망을 대상으로 강건성을 분석하였다. 실험 결과, ε 변화에 따른 강건성의 임계 구간을 식별하고 결정 경계 부근의 구체적 반례를 도출함으로써, 제안한 소프트웨어의 유효성을 확인하였다. 향후에는 표준 모델 형식으로부터 검증 문제를 자동 생성하는 인코딩 기법을 개발하고, 보다 대규모의 신경망에 대한 검증으로 확장할 계획이다.

[참고문헌]

[1] K. Eykholt et al., "Robust Physical-World Attacks on Deep Learning Visual Classification," CVPR, 2018.

[2] G. Katz et al., "Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks," CAV, 2017.

[3] VNN-COMP, https://sites.google.com/view/vnn2025

[4] K. Xu et al., "Automatic Perturbation Analysis for Scalable Certified Robustness and Beyond," NeurIPS, 2020.

[5] H. Duong et al., "NeuralSAT: A High-Performance Verification Tool for Deep Neural Networks," CAV, 2025.

[6] M. Davis et al., "A Machine Program for Theorem-Proving," Comm. ACM, vol. 5, no. 7, 1962.

[7] L. de Moura and N. Bjørner, "Z3: An Efficient SMT Solver," TACAS, 2008.

[8] A. Albarghouthi, "Introduction to Neural Network Verification," Found. Trends Program. Lang., vol. 7, no. 1-2, 2021.