김일성종합대학학보　　　　　　　　　　　　　　　　　　　　　　　주체110(2021)년
수 학　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　제67권　제1호

구간분석을 통한 심층신경망의 안정성검증의 한가지 방법

리철진, 최창일

경애하는 **김정은**동지께서는 다음과 같이 말씀하시였다.

《과학기술부문에서 첨단돌파전을 힘있게 벌려야 하겠습니다.》

선행연구[1]에서는 비선형활성함수를 가지는 신경망의 안정성을 정의하고 한계모형검사도구를 리용하여 안정성을 검증하기 위한 일반적인 한가지 방법을 제기하였다. 선행연구[2]에서는 화상분류를 진행하는 신경망에 대하여 주어진 화상과 주어진 잡음에 관한 안정성을 정의하고 술어론리를 리용하여 검증하는 방법을 제기하였다. 선행연구[3]에서는 일정한 조건을 만족시키고 ReLU활성함수를 리용하는 신경망에 대하여 선행연구[2]에서 정의된 안정성을 검증하는 방법과 도구를 제기하였다. 선행연구[4]에서는 주어진 입력모임에 관한 심층신경망의 안정성문제를 출력모임과 불안정모임의 사귐을 구하는 문제에로 귀착시키는 방법을 제기하였다.

론문에서는 활성함수에 대한 제한조건이 선행연구[3]에서보다 약화된 심층신경망의 안정성을 선행연구[4]에서 제기한 방법을 갱신하여 검증하는 한가지 방법을 제기하였다.

심층신경망 N을 다음과 같은 4원조$(L,\ W,\ b,\ Act)$로 표시한다.

① 층들의 모임 $L:\{L_k \mid L_k : n_k$개의 마디를 가진다. $0 \le k \le l\}$

② 무게결수행렬들의 모임

$$\boldsymbol{W}:\{\boldsymbol{W}^{\{k\}}=[w_1^{\{k\}},\ \cdots,\ w_{n_k}^{\{k\}}]^{\mathrm{T}},\ w_i^{\{k\}} \in \mathbf{R}^{n_{k-1}},\ 1 \le i \le n_k,\ 1 \le k \le l\}$$

③ 편향벡토르들의 모임

$$\boldsymbol{b}:\{\boldsymbol{b}^{\{k\}}=[b_1^{\{k\}},\ \cdots,\ b_{n_k}^{\{k\}}]^{\mathrm{T}},\ b_i^{\{k\}} \in \mathbf{R},\ 1 \le i \le n_k,\ 1 \le k \le l\}$$

④ 활성함수모임 $Act:\{\phi_k,\ 1 \le k \le l\}$

심층신경망 $N=(L,\ W,\ b,\ Act)$의 층 $L_k (1 \le k \le l)$가 층 L_{k-1}로부터 받아들이는 벡토르 $\boldsymbol{x}^{\{k-1\}}$에 대하여 이 층의 출력벡토르 $\boldsymbol{y}^{\{k\}}$는 다음과 같이 표시된다.

$$\boldsymbol{y}^{\{k\}}=\phi_k(\boldsymbol{W}^{\{k\}}\boldsymbol{x}^{\{k\}}+\boldsymbol{b}^{\{k\}})$$

특히 심층신경망 N의 출력층 L_l의 출력벡토르 $\boldsymbol{y}^{\{l\}}$은 다음과 같이 표시된다.

$$\boldsymbol{y}^{\{l\}}=\Phi(\boldsymbol{x}^{\{0\}}),\ \Phi(\boldsymbol{x}^{\{0\}}):=\hat{\phi}_l \circ \cdots \circ \hat{\phi}_1(\boldsymbol{x}^{\{0\}}),\ \hat{\phi}_k:=\phi_k(\boldsymbol{W}^{\{k\}}\boldsymbol{x}^{\{k\}}+\boldsymbol{b}^{\{k\}}) \tag{1}$$

정의 1[4] 심층신경망 $N=(L,\ W,\ b,\ Act)$과 주어진 입력벡토르들의 유한모임 $X \subseteq \mathbf{R}^{n_0}$에 대하여 식 (1)에 의해 얻어지는 모임

$$Y=\{\boldsymbol{y}^{\{l\}} \in \mathbf{R}^{n_l} \mid \boldsymbol{y}^{\{l\}}=\Phi(\boldsymbol{x}^{\{0\}}),\ \ \boldsymbol{x}^{\{0\}} \in X\}$$

를 X의 출력모임이라고 부른다. 또한 Φ를 N의 출력함수라고 부른다.

심층신경망의 검증에서는 특정한 입력벡토르들의 모임 $X \subseteq \mathbf{R}^{n_0}$에 대하여 그것의 출력모임의 원소들이 취해서는 안되는 모임 $S_X \subseteq \mathbf{R}^{n_l}$을 X에 관한 불안정모임이라고 부른다.

실례로 softmax활성함수를 리용하여 0부터 9까지의 필기체수자화상을 분류하는 심층신경망모형이 있다고 하자. 이 신경망에 대하여 수자 2를 나타내는 유한개의 화상들을 입력모임으로 할 때 출력벡토르에서 2에 대응하는 성분값이 0.5보다 작지 않으면 2로 정확히 분류된다.

이로부터 이 입력모임의 불안정모임 S_X는 $S_X = \{\boldsymbol{y} = [y_0, \cdots, y_9] \in \mathbf{R}^{10} \mid y_2 < 0.5\}$로 볼 수 있다.

일반적으로 불안정모임은 심층신경망의 구조와 입력모임에 따라 결정된다.

정의 2 신경망 $N = (L, W, b, Act)$와 입력벡토르들의 유한모임 $X \subseteq \mathbf{R}^{n_0}$, X의 불안정모임 S_X와 출력모임 Y에 대하여 $Y \cap S_X = \varnothing$이면 신경망 N은 입력 X에 대하여 안정하다고 말한다.

정의 3 $\mathbf{R}$우에서의 닫긴구간전부의 모임을 $\mathbf{IR}$로 표시하자. 함수 $\phi : \mathbf{R} \to \mathbf{R}$에 대하여 $[\phi]([x, x]) = \phi(x),\ x \in \mathbf{R}$를 만족시키는 함수 $[\phi] : \mathbf{IR} \to \mathbf{IR}$를 ϕ의 구간확장이라고 부른다. 마찬가지로 함수 $\Phi : \mathbf{R}^n \to \mathbf{R}^m\ (n,\ m \in \mathbf{N})$에 대하여

$$[\Phi]([x_1,\ x_1] \times \cdots \times [x_n,\ x_n]) = \Phi(\boldsymbol{x}),\quad \boldsymbol{x} = [x_1,\ \cdots,\ x_n] \in \mathbf{R}^n$$

이 성립하는 함수 $[\Phi] : \mathbf{IR}^n \to \mathbf{IR}^m$을 Φ의 구간확장이라고 부른다.

n차원벡토르들의 유한모임 $X \subseteq \mathbf{R}^n$에 대하여 X_i를 X의 원소들의 i째 성분들의 모임이라고 할 때 닫긴구간 $[\min(X_i),\ \max(X_i)],\ i = 1, \cdots, n$들의 직적을 $[X]$로 표시한다.

$[X] = [\underline{x}_1,\ \overline{x}_1] \times \cdots \times [\underline{x}_n,\ \overline{x}_n] \in \mathbf{IR}^n$에 대하여 $w([X]) = \max\limits_{i=1,\cdots,n} (\overline{x}_i - \underline{x}_i)$를 $[X]$의 너비라고 부른다.

정의 4 함수 $\Phi : \mathbf{R}^n \to \mathbf{R}^m\ (n,\ m \in \mathbf{N})$의 구간확장 $[\Phi] : \mathbf{IR}^n \to \mathbf{IR}^m$이 임의의 $[X_1],\ [X_2] \in \mathbf{IR}^n$에 대하여

$$[X_1] \subseteq [X_2] \Rightarrow [\Phi]([X_1]) \subseteq [\Phi]([X_2])$$

를 만족시키면 $[\Phi]$는 포함단조라고 부른다.

정의 4로부터 다음의 보조정리를 얻는다.

보조정리 1 함수 $\Phi : \mathbf{R}^n \to \mathbf{R}^m\ (n,\ m \in \mathbf{N})$의 구간확장 $[\Phi] : \mathbf{IR}^n \to \mathbf{IR}^m$이 포함단조일 때 임의의 n차원벡토르들의 유한모임 X에 대하여 다음의 성질이 만족된다.

$$\forall \boldsymbol{x} \in [X] \Rightarrow \Phi(\boldsymbol{x}) \in [\Phi]([X]) \tag{2}$$

증명 $[\Phi]$는 함수 Φ의 구간확장이므로 $\forall \boldsymbol{x} = [x_1, \cdots, x_n] \in [X]$에 대하여

$$\Phi(\boldsymbol{x}) = [\Phi]([x_1,\ x_1] \times \cdots \times [x_n,\ x_n])$$

이 성립한다. 또한 $[x_1,\ x_1] \times \cdots \times [x_n,\ x_n] \subset [X]$이며 $[\Phi]$는 포함단조인 구간확장이므로

$$[\Phi]([x_1,\ x_1] \times \cdots \times [x_n,\ x_n]) \subseteq [\Phi]([X])$$

가 성립하며 따라서 식 (2)가 만족된다.(증명끝)

보조정리 1로부터 다음의 보조정리를 얻는다.

보조정리 2 함수 $\Phi : \mathbf{R}^n \to \mathbf{R}^m\ (n,\ m \in \mathbf{N})$의 구간확장 $[\Phi] : \mathbf{IR}^n \to \mathbf{IR}^m$이 포함단조일 때 임의의 n차원벡토르들의 유한모임 X에 대하여 다음의 성질이 만족된다.

$$\Phi([X]) \subset [\Phi]([X]) \tag{3}$$

증명 보조정리 1로부터 $\forall \boldsymbol{x}\in[X]$에 대하여 $\Phi(\boldsymbol{x})\in[\Phi]([X])$가 성립하며 따라서 식 (3)이 만족된다.(증명끝)

보조정리 2로부터 신경망 $N=(L,\ W,\ b,\ Act)$의 입력벡토르들의 유한모임 X와 그것의 불안정모임 S_X에 대하여 출력함수 Φ의 구간확장 $[\Phi]$가 포함단조이고 $[\Phi]([X])\cap S_X=\varnothing$이면 N은 X에서 안정하다는것을 알수 있다.

출력함수 Φ의 포함단조인 구간확장 $[\Phi]$를 구하기 위하여 신경망의 활성함수에 대한 한가지 가정을 제기한다.

가정 신경망 $N=(L,\ W,\ b,\ Act)$의 활성함수 $\phi_k(k=1,\ \cdots,\ l)$는 단조증가함수이다. 즉

$$\forall x_1,\ x_2\in\mathbf{R},\ x_1\le x_2\Rightarrow\phi_k(x_1)\le\phi_k(x_2)$$

심층신경망에서 리용되는 ReLU함수, tanh함수, 로지스틱함수, sigmoid함수 등 대부분의 활성함수들은 모두 가정을 만족시킨다.

가정을 만족시키는 활성함수를 리용하는 신경망에 대하여 출력함수의 포함단조인 구간확장을 구성하는 한가지 방법을 제기한다.

정리 가정을 만족시키는 신경망 $N=(L,\ W,\ b,\ Act)$와 N의 입력벡토르들의 유한모임 X에 대하여 다음과 같이 구성되는 함수 $[\Phi]$는 출력함수 Φ의 구간확장이며 포함단조이다.

$$[\Phi]([X])=[\hat{\phi}_l]\circ\cdots\circ[\hat{\phi}_1]([X])\tag{4}$$

여기서

$$[\hat{\phi}_k]([X^{\{k\}}]):=[\phi_k](\boldsymbol{W}^{\{k\}}[X^{\{k\}}]+\boldsymbol{b}^{\{k\}})=[Y^{\{k\}}]=[\underline{y}_1^{\{k\}},\ \overline{y}_1^{\{k\}}]\times\cdots\times[\underline{y}_{n_k}^{\{k\}},\ \overline{y}_{n_k}^{\{k\}}]\tag{5}$$

$$\left.\begin{array}{l}\underline{y}_i^{\{k\}}:=\phi_k\left(\displaystyle\sum_{j=1}^{n_{k-1}}\underline{p}_{ij}^{\{k\}}+b_i^{\{k\}}\right),\ \underline{p}_{ij}^{\{k\}}:=\begin{cases}w_{ij}^{\{k\}}\underline{x}_j^{\{k\}},\ w_{ij}^{\{k\}}\ge0\\ w_{ij}^{\{k\}}\overline{x}_j^{\{k\}},\ w_{ij}^{\{k\}}<0\end{cases}\\ \overline{y}_i^{\{k\}}:=\phi_k\left(\displaystyle\sum_{j=1}^{n_{k-1}}\overline{p}_{ij}^{\{k\}}+b_i^{\{k\}}\right),\ \overline{p}_{ij}^{\{k\}}:=\begin{cases}w_{ij}^{\{k\}}\overline{x}_j^{\{k\}},\ w_{ij}^{\{k\}}\ge0\\ w_{ij}^{\{k\}}\underline{x}_j^{\{k\}},\ w_{ij}^{\{k\}}<0\end{cases}\end{array}\right.\quad(i=1,\ \cdots,\ n_k)\tag{6}$$

이며 $[X^{\{k\}}]=[Y^{\{k-1\}}],\ [X^{\{1\}}]=[X]$이다.

증명 먼저 $[\Phi]$가 출력함수 Φ의 구간확장이라는것을 밝히자.

임의의 입력벡토르 $\boldsymbol{x}=[x_1,\ \cdots,\ x_{n_0}]\in\mathbf{R}^{n_0}$에 대하여 $[x_1,\ x_1]\times\cdots\times[x_{n_0},\ x_{n_0}]$을 $[\boldsymbol{x}]$로 표시하면

$$[\hat{\phi}_1]([\boldsymbol{x}])=[\underline{y}_1^{\{1\}},\ \overline{y}_1^{\{1\}}]\times\cdots\times[\underline{y}_{n_1}^{\{1\}},\ \overline{y}_{n_1}^{\{1\}}]$$

$$\underline{y}_i^{\{1\}}=\overline{y}_i^{\{1\}}=\phi\left(\sum_{j=1}^{n_0}w_{ij}^{\{1\}}x_j+b_i^{\{1\}}\right)$$

로서 $[\hat{\phi}_1]([x_1,\ x_1]\times\cdots\times[x_{n_0},\ x_{n_0}])$은 $\hat{\phi}_1(\boldsymbol{x})$와 같다. 마찬가지로

$$[\Phi]([\boldsymbol{x}])=[\hat{\phi}_l]\circ\cdots\circ[\hat{\phi}_1]([\boldsymbol{x}])=\hat{\phi}_l\circ\cdots\circ\hat{\phi}_1(\boldsymbol{x})=\Phi(\boldsymbol{x})$$

가 성립하며 따라서 $[\Phi]$는 출력함수 Φ의 구간확장이다.

다음으로 $[\Phi]$가 포함단조이라는것을 밝히자.

$[X_1]\subseteq[X_2]$인 $[X_m]=[\underline{x}_{m,1},\ \overline{x}_{m,1}]\times\cdots\times[\underline{x}_{m,n_0},\ \overline{x}_{m,n_0}],\ m=1,\ 2$가 있다고 하자. 그러면

$$\underline{x}_{2,j}\le\underline{x}_{1,j}\le\overline{x}_{1,j}\le\overline{x}_{2,j},\ \ j=1,\ \cdots,\ n_0 \tag{7}$$

이 성립한다. $[X_m],\ m=1,\ 2$에 $[\hat{\phi}_1]$을 적용한 경우를 보기로 하자.

$$[\hat{\phi}_1]([X_m])=[\underline{y}^{\{1\}}_{m,1},\ \overline{y}^{\{1\}}_{m,1}]\times\cdots\times[\underline{y}^{\{1\}}_{m,n_1},\ \overline{y}^{\{1\}}_{m,n_1}],\ m=1,\ 2$$

$$\left.\begin{aligned}\underline{y}^{\{1\}}_{m,i}&=\phi_1(\sum_{j=1}^{n_0}\underline{p}^{\{1\}}_{m,ij}+b^{\{1\}}_i),\ \ \underline{p}^{\{1\}}_{m,ij}=\begin{cases}w^{\{1\}}_{ij}\underline{x}_{m,j},\ w^{\{1\}}_{ij}\ge 0\\ w^{\{1\}}_{ij}\overline{x}_{m,j},\ w^{\{1\}}_{ij}<0\end{cases}\\ \overline{y}^{\{1\}}_{m,i}&=\phi_1\left(\sum_{j=1}^{n_0}\overline{p}^{\{1\}}_{m,ij}+b^{\{1\}}_i\right),\ \ \overline{p}^{\{1\}}_{m,ij}=\begin{cases}w^{\{1\}}_{ij}\overline{x}_{m,j},\ w^{\{1\}}_{ij}\ge 0\\ w^{\{1\}}_{ij}\underline{x}_{m,j},\ w^{\{1\}}_{ij}<0\end{cases}\end{aligned}\right.\quad (i=1,\ \cdots,\ n_1)$$

식 (7)로부터 $\underline{p}^{\{1\}}_{1,ij}\ge\underline{p}^{\{1\}}_{2,ij},\ \overline{p}^{\{1\}}_{2,ij}\ge\overline{p}^{\{1\}}_{1,ij}\ (i=1,\ \cdots,\ n_1,\ j=1,\ \cdots,\ n_0)$이 성립한다는것을 쉽게 알수 있다. 또한 가정에 의하여 활성함수 $\phi_k(k=1,\ \cdots,\ l)$는 단조증가함수이므로

$$\underline{y}^{\{1\}}_{1,i}=\phi_1\left(\sum_{j=1}^{n_0}\underline{p}^{\{1\}}_{1,ij}+b^{\{1\}}_i\right)\ge\phi_1\left(\sum_{j=1}^{n_0}\underline{p}^{\{1\}}_{2,ij}+b^{\{1\}}_i\right)=\underline{y}^{\{1\}}_{2,i}$$

$$\overline{y}^{\{1\}}_{1,i}=\phi_1\left(\sum_{j=1}^{n_0}\overline{p}^{\{1\}}_{1,ij}+b^{\{1\}}_i\right)\le\phi_1\left(\sum_{j=1}^{n_0}\overline{p}^{\{1\}}_{2,ij}+b^{\{1\}}_i\right)=\overline{y}^{\{1\}}_{2,i}$$

이 성립하며 따라서 $[\underline{y}^{\{1\}}_{1,i},\ \overline{y}^{\{1\}}_{1,i}]\subseteq[\underline{y}^{\{1\}}_{2,i},\ \overline{y}^{\{1\}}_{2,i}],\ i=1,\ \cdots,\ n_1$로서 $[\hat{\phi}_1]([X_1])\subseteq[\hat{\phi}_1]([X_2])$도 성립한다. 즉 $[\hat{\phi}_1]$은 포함단조이다.

마찬가지로 $[\hat{\phi}_l]\circ\cdots\circ[\hat{\phi}_1]([X_1])\subseteq[\hat{\phi}_l]\circ\cdots\circ[\hat{\phi}_1]([X_2])$도 성립하며 $[\Phi]$가 포함단조라는것이 증명된다.(증명끝)

정리 1과 보조정리 2로부터 다음과 같은 따름을 얻게 된다.

따름 가정을 만족시키는 신경망 $N=(L,\ \boldsymbol{W},\ \boldsymbol{b},\ Act)$와 N의 입력벡토르들의 유한모임 X, 그것의 불안정모임 S_X, 식 (4)-(6)과 같이 구성된 출력함수 Φ의 구간확장 $[\Phi]$에 대하여 $[\Phi]([X])\cap S_X=\varnothing$이면 N은 X에 대하여 안정하다.

우의 따름에 기초하여 가정을 만족시키는 신경망 $N=(L,\ \boldsymbol{W},\ \boldsymbol{b},\ Act)$와 N의 입력벡토르들의 유한모임 X와 그것의 불안정모임 S_X가 주어졌을 때 안정성을 검증하는 알고리듬을 제기한다.

알고리듬

입력: 가정을 만족시키는 신경망 $N=(L,\ \boldsymbol{W},\ \boldsymbol{b},\ Act)$와 입력모임 $X\subseteq\mathbf{R}^{n_0}$, X의 불안정모임 $S_X\subseteq\mathbf{R}^{n_l}$, 턱값 $\varepsilon>0$

출력: 모임 *Unsafe*

단계 1: $M\leftarrow\{[X]\}$, $Unsafe\leftarrow\{\}$

단계 2: M이 빈모임이면 단계 7로 넘어간다. 아니면 단계 3에로 넘어간다.

단계 3: M에서 한 원소 $[X]$를 선택한다. 동시에 M에서 그 원소를 제거한다.

단계 4: $[\Phi]([X])\cap S_X$가 빈모임이면 단계 2로 넘어간다. 아니면 단계 5로 넘어간다.

단계 5: $w([X])>\varepsilon$ 이면 단계 6에로 넘어간다. 아니면 *Unsafe* 에 $[X]$ 를 추가하고 단계 2로 넘어간다.

단계 6: $[X]=[\underline{x}_1,\ \overline{x}_1]\times\cdots\times[\underline{x}_{n_0},\ \overline{x}_{n_0}]$ 에서 $\overline{x}_i-\underline{x}_i=w([X])$ 인 한 닫긴구간을 $[\underline{x}_i,\ \underline{x}_i+w([X])/2]$, $[\underline{x}_i+w([X])/2,\ \overline{x}_i]$ 로 분할하여 $[X_1]$, $[X_2]$ 를 구성하고 모임 M 에 추가한 다음 단계 2로 넘어간다.

단계 7: *Unsafe* 를 출력한다.

모임 *Unsafe* 가 빈모임이면 신경망 N 은 X 에 대하여 안정하다.

참 고 문 헌

[1] L. Pulina et al.; AI Communications, 25, 2, 117, 2012.
[2] X. Huang et al.; CAV, LNCS, 10426, 3, 2017.
[3] G. Katz et al.; CAV, LNCS, 10426, 97, 2017.
[4] W. Xiang et al.; IEEE Transactions on Neural Network and Learning Systems, 22, 5777, 2018.

주체109(2020)년 9월 5일 원고접수

Safety Verification of Deep Neural Network Using Interval Analysis

Ri Chol Jin, Choe Chang Il

A method based on interval analysis for safety verification of deep neural network is presented in this paper.

Keywords: deep neural network, safety verification, interval analysis