Formal verification of neural networks

Dmitry Kondratyev

Ph.D. in Mathematics and Computer Science
Researcher
The Artificial Intelligence Research Center of Novosibirsk State University

Problem: how to achieve trustworthy artificial intelligence

Only formal verification guarantees software correctness.

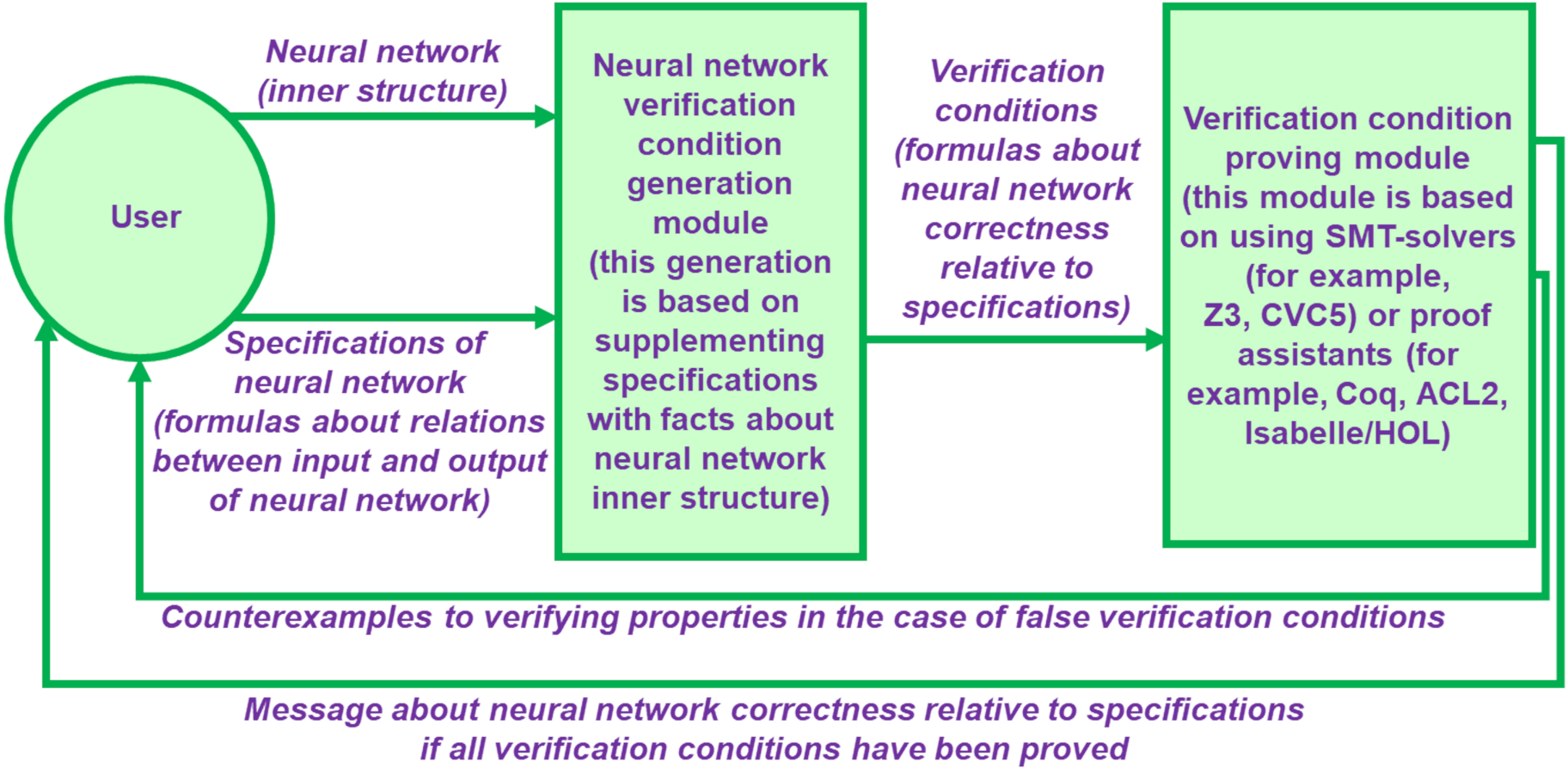
Stages of **deductive verification** (important kind of formal verification):

1. Defining program specifications. Program and its specification can be represented as Hoare triple: $\{P\}\, S\, \{Q\}$, where P — precondition (formula about properties of input data), S — program, Q — postcondition (formula about properties of output data).
2. Generating verification conditions. It is formulas about correctness of program relative to its specifications. Generation of verification condition is based on moving from conclusions to premises of inference rules of Hoare logic. Let us consider scheme of such rules:

$$\frac{\{P_1\}\, S_1\, \{Q_1\},\ \ldots\ \{P_n\}\, S_n\, \{Q_n\},\ \gamma_1,\ \ldots\ \gamma_m}{\{P\}\, S\, \{Q\}}$$

   where $\{P_1\}\, S_1\, \{Q_1\},\ \ldots$ — premises (Hoare triples), $\gamma_1,\ \ldots$ — premises (verification conditions) and $\{P\}\, S\, \{Q\}$ — conclusion (Hoare triple)
3. Proving verification conditions. If all verification conditions are valid then program is correct relative to its specifications. SMT-solvers and interactive theorem provers can be applied at this stage.
4. Localizing errors (counterexamples to invalid verification conditions)

Scheme of neural network formal verification



The paper about deductive verification of neural networks

Let us consider review of the following paper:
Xie X., Kersting K., Neider D. Neuro-Symbolic Verification of Deep Neural Networks. Proc. 31 Int. Joint Conf. on Artificial Intelligence. 2022. pp. 3622–3628. DOI: https://doi.org/10.24963/ijcai.2022/503

A deep neural network $f : \mathbb{R}^m \to \mathbb{R}^n$ can be considered as an extended graph $G_f = (V, V_I, V_O, E, \alpha)$ where V is a finite set of vertices (i.e., neurons), V_I are the input neurons, V_O are the output neurons, $E \subseteq V \times \mathbb{R} \times V$ is a weighted, directed edge relation and α is a mapping that assigns an activation function to each neuron in $V \setminus V_I$.

Hoare triple for network: $\{\varphi_{pre}(\vec{x})\}\ \varphi_{assign}(\vec{x}, \vec{y})\ \{\varphi_{post}(\vec{x}, \vec{y})\}$, where

- $f : \mathbb{R}^m \to \mathbb{R}^n$ — function representation of neural networks f,
- $\vec{x} = (x_1, \ldots, x_m)$ — vector representing the input values of f,
- $\vec{y} = (y_1, \ldots, y_n)$ — vector representing the output values of f,
- the expression $\varphi_{assign}(\vec{x}, \vec{y}) := \vec{y} \leftarrow f(\vec{x})$ stores the result of the computation $f(\vec{x})$ in the variable $\vec{y}$,
- $G_f = (V, V_I, V_O, E, \alpha)$ — graph representation of neural network f,
- φ_{pre} — formula over $\vec{x}_1, \ldots, \vec{x}_\ell$ representing precondition,
- φ_{post} — formula over $\vec{x}_1, \ldots, \vec{x}_\ell$, $\vec{y}_1, \ldots \vec{y}_k$ representing postcondition.

The paper about deductive verification of neural networks

Let variable X_v correspond to output value of neuron $v \in V$. Then formula φ_v corresponds to neuron $v \in V$ (except input neurons):

$$\varphi_v \;:=\; X_v = \alpha(v)\left(\sum_{(v',w,v)\in E} w \cdot X_{v'}\right)$$

Let define φ_v using auxiliary Y_v in the case of ReLU activation function:

$$\left(Y_v = \textstyle\sum_{(v',w,v)\in E} w \cdot X_{v'}\right) \wedge \Big(\big(Y_v \le 0 \to X_v = 0\big) \wedge \big(Y_v > 0 \to X_v = Y_v\big)\Big)$$

Thus, the following formula corresponds to whole network: $\varphi_f \;:=\; \bigwedge_{v\in V\setminus V_I} \varphi_v$
Inference rule for neural network (formal semantics of neural network):

$$\frac{\Big(\varphi_{pre}[\vec{x}\,/\,\vec{X}_{V_I}] \wedge \varphi_f\Big) \to \varphi_{post}[\vec{x}\,/\,\vec{X}_{V_I}, \vec{y}\,/\,\vec{X}_{V_O}]}{\Big\{\varphi_{pre}(\vec{x})\Big\}\,\vec{y} \leftarrow f(\vec{x})\,\Big\{\varphi_{post}(\vec{x},\vec{y})\Big\}}$$

where $\varphi[\vec{z}_1\,/\,\vec{z}_2]$ is used to denote the formula resulting from the substitution of the vector of variables z_1 by z_2 in φ.
Verification condition: $\big(\varphi_{pre}[\vec{x}\,/\,\vec{X}_{V_I}] \wedge \varphi_f\big) \to \varphi_{post}[\vec{x}\,/\,\vec{X}_{V_I}, \vec{y}\,/\,\vec{X}_{V_O}]$
If this verification condition is valid then neural network is correct relative to its specifications. It allows achieving trustworthy artificial intelligence.

Towards smart vehicles: review of the paper about robot

Amir G., Corsi D., Yerushalmi R., Marzari L., Harel D., Farinelli A., Katz G. Verifying Learning-Based Robotic Navigation Systems // Lecture Notes in Computer Science. 2023. Volume 13993. pp. 607–627.

The robot moves towards the target (the angle and distance are known). The robot is controlled by the following commands: forward, turn right by 30^0 and turn left by 30^0. The robot has 7 lidars to measure the distance to obstacles in 7 directions. The neural network receives data about the target and obstacles as input and generates a command to the robot.

Importance of formal verification of neural networks

780 neural networks were trained using reinforcement learning methods to control such a robot.

780 neural networks were verified. Counterexamples to at least one property were found for 778 neural networks. And only for 2 neural networks was proven that all properties were satisfied (they were trained using the Actor-Critic method).

In summary, reinforcement learning methods do not allow obtaining neural networks that are correctly constructed.

Neural network formal verification in Russia

Publications about neural network formal verification in Russia:

- Полевой А.В., Корухова Ю.С. Об одном подходе к формальной верификации нейросетевых моделей // Научный сервис в сети Интернет: труды XXVI Всероссийской научной конференции (23–25 сентября 2024 г., онлайн). – М.: ИПМ им. М.В.Келдыша, 2024. – С. 223–235. DOI: https://doi.org/10.20948/abrau-2024-11
- Строева Е.Н., Тонких А.А. Методы формальной верификации искусственных нейронных сетей: обзор существующих подходов // International Journal of Open Information Technologies. 2022. Т. 10. № 10. С. 21–29.
- Селевенко Р.М., Строева Е.Н. Исследование и разработка алгоритма формальной верификации и метрики оценки качества на основе методов понижения размерности ИНС // International Journal of Open Information Technologies. 2024. Т. 12. № 6. С. 14–26.

Formal verification group

head, senior researcher
Ph.D. in Mathematics and Computer Science
Anureev I.S.

senior researcher
Ph.D. in Mathematics and Computer Science
Garanina N.O.

researcher
Ph.D. in Mathematics and Computer Science
Kondratyev D.A.

The Artificial Intelligence Research Center of Novosibirsk State University

The A.P. Ershov Institute of Informatics Systems SB RAS

Proposed ideas of comprehensive approach to formal verification of neural networks

Ideas of comprehensive approach to formal verification of neural networks have been proposed in the Artificial Intelligence Research Center of Novosibirsk State University.

Proposed ideas are based on the comprehensive approach to deductive verification of C programs developed in the A.P. Ershov Institute of Informatics Systems SB RAS. The following paper contains descriptions of this approach and its implementation in the C-lightVer system:

Kondratyev D.A., Nepomniaschy V.A. Automation of C Program Deductive Verification without Using Loop Invariants. Programming and Computer Software. 2022. Volume 48. Issue 5. pp. 331–346. DOI: `https://doi.org/10.1134/S036176882205005X`

Comprehensive approach to formal verification of neural networks will include:

- Language for defining specifications of neural networks.
- Methods of defining models of neural networks.
- Set of inference rules for generating verification conditions of neural networks (formal semantics of neural networks).

Formal verification of neural networks

Dmitry Kondratyev

Ph.D. in Mathematics and Computer Science
Researcher
The Artificial Intelligence Research Center of Novosibirsk State University