神经网络形式化验证算法调研

智能系统的安全可信性问题不只是在人工智能或者机器学习领域得到关注，也 引起了形式化方法领域的关注。传统上,形式化方法分析和验证主要针对软硬件系统。尽管以神经网络为代表的智能系统也可以实现为软件或者硬件，一个显著的不同在于,传统形式化方法研究的软硬件系统一般有严格的逻辑描述或者结构化描述，也就是符号系统。而智能网络缺乏逻辑结构，神经网络是通过大量的神经元之间的连接构造起 来的系统，并通过优化算法来学习系统参数。从符号到亚符号系统的变迁使得形式化方法领域半个世纪来发展的大量技术并不能直接被用于智能系统的可信性研究。

下面我们从形式化方法的角度出发观察过去几年基于形式化方法的智能系统安全可信研究的进展。

# 1 方法分类

从方法所运用到的技术来分，大致可以分为三类：

1. 以可达性为基础方法
2. 以优化为基础的方法
3. 以搜索为基础的方法

# 2 基于可达性的方法

可达性描述：

在给定某个神经网络和输入集合的前提下，其对应的输出值会有一个可达集，称之为输出可达集。此类方法的原理是，一方面可以计算出这个输出可达集的某个近似估计，另一方面也可以检查这一输出可达集与某些安全规格（specification）的非（negation）的交集是否为空。

相关论文及工具：

【ExactReach】Xiang, W., H.-D. Tran, and T. T. Johnson. 2017. “Reachable set computation and

safety verification for neural networks with ReLU activations”. ArXiv. (1712.08163).

【Ai2】Gehr, T., M. Mirman, D. Drashsler-Cohen, P. Tsankov, S. Chaudhuri, and M. Vechev.

2018. “Ai2: Safety and robustness certification of neural networks with abstract

interpretation”. In: IEEE Symposium on Security and Privacy (SP).

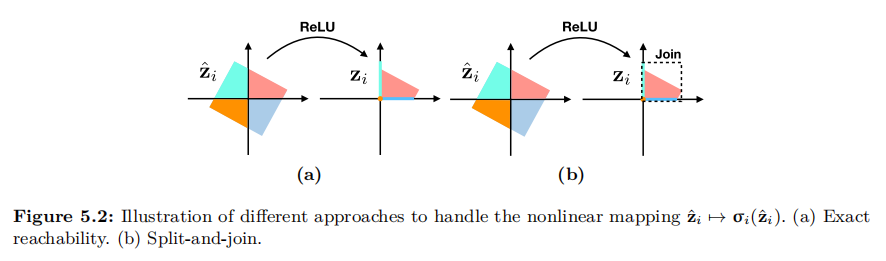
【MaxSens】Xiang, W., H. Tran, and T. T. Johnson. 2018a. “Output Reachable Set Estimation and

Verification for Multilayer Neural Networks”. IEEE Transactions on Neural Networks

and Learning Systems. 29(11): 5777–5783.

## 2.1 ExactReach

ExactReach 采用了几何计算解决了基于可达性的神经网络验证问题。ExactReach对具有线性或ReLU激活函数的网络执行精确的可达性分析。对于任何ReLU函数，如果输入集是多面体的并集，则输出可达集也是多面体的并集，如图5.2a所示。该方法可以用H-多面体或星集实现。



特点：对于一个输入多面体，ExactReach方法每经过一个隐藏层就会生成2ki个多边形（ki是第i层的约束数量）。因此多边形的数量随着深度的增加呈指数级增长。虽然空集可以在这个过程中被删除，但跟踪大型神经网络的精确可达集仍然显得效率低下。

解决这个问题的方法是并行计算个别几何对象，或者开发新的数据结构来处理线性和非线性映射。例如之后的星集。

## 2.2 Ai2

就像ExactReach中提到的，在许多情况下，精确可达集是难以处理的，在Ai2中，使用了可达集的估计，它满足原精确可达集是估计可达集的子集。

Ai2使用抽象域来近似每一层的可达集，该集合由捕获某些几何形状的一组逻辑公式表示，例如半空间多面体、超矩形、全对称多面体、星集、半空间等。抽象域的选择需要在精度和可扩展性之间取得平衡。 例如，在ExactReach中使用的多面体是精确的，但可扩展性差，而在MaxSens中使用的超矩形是可缩放的，但太松散。

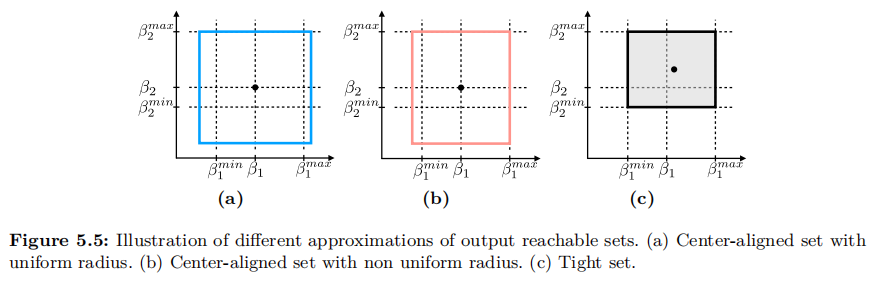
特点：Ai2的原始实现方式使用了zonotopes，即中心对称的凸封闭多面体，它比多面体更具可伸缩性，比超矩形更紧密。相比于其他两种方法，这种方法显得更加均衡。

## 2.3 MaxSens

MaxSens是一种使用过近似的可达性方法。它适用于具有单调激活函数以及低维输入输出的网络。MaxSens的关键思想是对输入空间进行网格化，并计算每个网格单元的可达集。 网格单元越精细，则过逼近度越小。 计算不同小区的可达集可以并行完成。

有三种方法来定义过近似：

1. 均匀半径的中心对齐集合，对应a) 图
2. 非均匀半径的中心对齐集合，对应b) 图
3. 紧密集合，对应c) 图



特点：与其他可达性方法相比，MaxSens的优势在于，在逐层传播期间，几何对象的数量不会增加。 超矩形的总数仅取决于初始分区。尽管在逐层传播期间超矩形的数目不会增加，但是相对于层数，过近似造成的误差将迅速累积。对于具有许多输入节点的网络，在严格估计时，初始分区上的超矩形数量可能会过大。否则，使用稀疏分区的计算将过于保守。

# 3 基于搜索的方法

基于搜索的方法一般会结合可达性或优化来一起使用，从而降低整体的时间复杂度，常见的基于搜索的方法有：分支定界法、SAT、SMT等。

相关论文：

## Sherlock

Sherlock一次只能计算出一个输出节点的上下界，通过结合局部搜索和全局搜索来计算输出节点的上下界。在全局搜索中，它求解一个线性规划问题，以检查是否可以优化当前的局部最优边界。在局部搜索中，求解器从一个样本点开始，在f的相同线段上找到一个与x处激活模式相同且更好的可行解。为了找到全局最优边界，求解器迭代执行局部搜索和全局搜索。

