The Third International Verification of Neural Networks Competition (VNN-COMP 2022): Summary and Results

Stanley Bak[*], Changliu Liu[†], Taylor Johnson[‡]

Abstract

This report summarizes the third International Verification of Neural Networks Competition (VNN-COMP 2021), held as a part of the 5th Workshop on Formal Methods for ML-Enabled Autonomous Systems (FOMLAS) that was collocated with the 34th International Conference on Computer-Aided Verification (CAV). The goal of the competition is to provide an objective comparison of the state-of-the-art methods in neural network verification, in terms of scalability and speed. Along this line, we used standard formats (ONNX for neural networks and VNNLIB for specifications), standard hardware (all tools are run by the organizers on AWS), and tool parameters provided by the tool authors. This report summarizes the rules, benchmarks, participating tools, results, and lessons learned from this competition.

1 Total

Table 1: Overall Score

#	Tool	Score
1	α,β-CROWN	1274.8
2	MN BaB	980.6
3	Verinet	893.8
4	Nnenum	533.9
5	Cgdtest	406.7
6	Peregrinn	399.4
7	Marabou	380.8
8	Debona	222.9
9	Fastbatllnn	100.0
10	Verapak	98.3
11	Averinn	29.1

[*]S. Bak is with Stony Brook University, stanley.bak@stonybrook.edu.
[†]C. Liu is with Carnegie Mellon University, cliu6@andrew.cmu.edu.
[‡]T. Johnson is with Vanderbilt University, taylor.johnson@vanderbilt.edu.

2 Scored

Table 2: Benchmark `carvana-unet-2022`

#	Tool	Verified	Falsified	Fastest	Score	Percent
1	α,β-CROWN	39	0	39	468	100.0%
2	MN BaB	19	0	0	209	44.7%
3	Verinet	3	0	0	30	6.4%

Table 3: Benchmark `cifar100-tinyimagenet-resnet`

#	Tool	Verified	Falsified	Fastest	Score	Percent
1	α,β-CROWN	69	0	56	813	100.0%
2	Cgdtest	95	0	28	725	89.2%
3	MN BaB	60	3	10	674	82.9%
4	Verinet	48	3	6	541	66.5%

Table 4: Benchmark `cifar-biasfield`

#	Tool	Verified	Falsified	Fastest	Score	Percent
1	α,β-CROWN	69	1	1	735	100.0%
2	Cgdtest	71	0	55	732	99.6%
3	Verinet	69	1	0	721	98.1%
4	Verapak	71	0	0	635	86.4%
5	MN BaB	36	1	17	404	55.0%
6	Marabou	27	0	0	270	36.7%
7	Nnenum	4	0	0	43	5.9%

Table 5: Benchmark `collins-rul-cnn`

#	Tool	Verified	Falsified	Fastest	Score	Percent
1	Nnenum	16	45	57	725	100.0%
2	MN BaB	16	44	57	715	98.6%
3	α,β-CROWN	15	45	56	612	84.4%
4	Verinet	16	43	0	590	81.4%
5	Peregrinn	14	42	0	560	77.2%
6	Cgdtest	1	42	43	-984	0%

Table 6: Benchmark `mnist-fc`

#	Tool	Verified	Falsified	Fastest	Score	Percent
1	α,β-CROWN	66	18	53	963	100.0%
2	Verinet	53	18	50	817	84.8%
3	MN BaB	53	18	47	804	83.5%
4	Debona	48	18	38	737	76.5%
5	Nnenum	48	11	29	648	67.3%
6	Marabou	44	16	0	600	62.3%
7	Peregrinn	27	11	7	394	40.9%
8	Cgdtest	66	3	23	241	25.0%
9	Verapak	40	2	42	104	10.8%

Table 7: Benchmark `nn4sys`

#	Tool	Verified	Falsified	Fastest	Score	Percent
1	α,β-CROWN	152	0	132	1791	100.0%
2	Verinet	57	0	43	670	37.4%
3	MN BaB	42	0	19	467	26.1%
4	Peregrinn	24	0	22	284	15.9%
5	Nnenum	23	0	8	246	13.7%
6	Debona	2	0	2	24	1.3%
7	Cgdtest	2	0	2	-176	0%

Table 8: Benchmark `oval21`

#	Tool	Verified	Falsified	Fastest	Score	Percent
1	α,β-CROWN	25	1	10	291	100.0%
2	MN BaB	19	1	2	205	70.4%
3	Verinet	17	1	1	189	64.9%
4	Marabou	19	0	17	125	43.0%
5	Nnenum	3	1	0	40	13.7%
6	Peregrinn	1	0	0	10	3.4%
7	Cgdtest	11	0	1	-580	0%

Table 9: Benchmark `reach-prob-density`

#	Tool	Verified	Falsified	Fastest	Score	Percent
1	Nnenum	22	14	21	410	100.0%
2	α,β-CROWN	22	14	23	406	99.0%
3	Verinet	22	14	10	383	93.4%
4	MN BaB	22	12	14	368	89.8%
5	Marabou	17	14	12	334	81.5%
6	Peregrinn	18	14	2	324	79.0%
7	Cgdtest	0	5	5	60	14.6%
8	Debona	0	2	2	24	5.9%

Table 10: Benchmark `rl-benchmarks`

#	Tool	Verified	Falsified	Fastest	Score	Percent
1	α,β-CROWN	193	103	296	3552	100.0%
2	Verinet	193	103	292	3547	99.9%
3	MN BaB	193	103	288	3536	99.5%
4	Nnenum	191	103	282	3504	98.6%
5	Peregrinn	193	103	271	3502	98.6%
6	Marabou	191	103	278	3496	98.4%
7	Debona	153	99	240	3000	84.5%
8	Averinn	92	8	16	1032	29.1%
9	Cgdtest	10	24	29	98	2.8%
10	Verapak	0	4	0	40	1.1%

Table 11: Benchmark `sri-resnet-a`

#	Tool	Verified	Falsified	Fastest	Score	Percent
1	α,β-CROWN	20	12	7	356	100.0%
2	Cgdtest	26	6	14	352	98.9%
3	MN BaB	18	12	19	343	96.3%
4	Verinet	12	12	4	248	69.7%
5	Marabou	3	0	0	30	8.4%

Table 12: Benchmark `sri-resnet-b`

#	Tool	Verified	Falsified	Fastest	Score	Percent
1	α,β-CROWN	28	11	8	432	100.0%
2	MN BaB	27	11	24	431	99.8%
3	Cgdtest	22	10	0	329	76.2%
4	Verinet	20	11	4	319	73.8%
5	Marabou	61	0	36	-411	0%

Table 13: Benchmark `tllverifybench`

#	Tool	Verified	Falsified	Fastest	Score	Percent
1	Fastbatllnn	11	21	32	384	100.0%
2	MN BaB	11	21	21	364	94.8%
3	α,β-CROWN	11	21	11	351	91.4%
4	Peregrinn	10	21	7	324	84.4%
5	Verinet	11	21	0	320	83.3%
6	Nnenum	1	21	10	240	62.5%
7	Debona	0	19	10	210	54.7%
8	Marabou	4	15	2	194	50.5%
9	Cgdtest	0	9	6	2	0.5%

Table 14: Benchmark `vggnet16-2022`

#	Tool	Verified	Falsified	Fastest	Score	Percent
1	α,β-CROWN	14	1	11	176	100.0%
2	Nnenum	11	1	0	127	72.2%
3	MN BaB	5	1	4	69	39.2%
4	Verinet	5	1	0	60	34.1%
5	Cgdtest	0	2	1	-378	0%

3 Unscored

Table 15: Benchmark `acasxu`

#	Tool	Verified	Falsified	Fastest	Score	Percent
1	Nnenum	139	47	174	2218	100.0%
2	α,β-CROWN	139	43	56	1685	76.0%
3	Cgdtest	85	30	115	680	30.7%

Table 16: Benchmark `cifar2020`

#	Tool	Verified	Falsified	Fastest	Score	Percent
1	α,β-CROWN	148	38	176	1822	100.0%
2	MN BaB	93	28	14	1313	72.1%
3	Nnenum	66	19	0	850	46.7%
4	Cgdtest	86	30	6	499	27.4%
5	Verapak	0	15	0	150	8.2%
6	Verinet	91	47	7	-3868	0%
7	Marabou	4	0	0	-60	0%

4 Stats