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 Score

Table 1: Overall Score

#	Tool	Score
1	α,β-CROWN	1274.9
2	MN BaB	1017.3
3	Verinet	892.5
4	Nnenum	534.0
5	Cgdtest	406.4
6	Peregrinn	399.0
7	Marabou	380.6
8	Debona	222.9
9	Fastbatllnn	100.0
10	Verapak	98.2
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 Benchmarks

Table 2: Benchmark `carvana-unet-2022`

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

Table 3: Benchmark `cifar100-tinyimagenet-resnet`

#	Tool	Verified	Falsified	Fastest	Penalty	Score	Percent
1	α,β-CROWN	69	0	56	0	813	100.0%
2	Cgdtest	95	0	28	3	725	89.2%
3	MN BaB	60	3	10	0	674	82.9%
4	Verinet	48	3	6	0	540	66.4%

Table 4: Benchmark `cifar-biasfield`

#	Tool	Verified	Falsified	Fastest	Penalty	Score	Percent
1	α,β-CROWN	69	1	1	0	736	100.0%
2	Cgdtest	71	0	55	1	731	99.3%
3	Verinet	69	1	0	0	721	98.0%
4	Verapak	71	0	0	1	635	86.3%
5	MN BaB	36	1	17	0	404	54.9%
6	Marabou	27	0	0	0	270	36.7%
7	Nnenum	4	0	0	0	43	5.8%

Table 5: Benchmark `collins-rul-cnn`

#	Tool	Verified	Falsified	Fastest	Penalty	Score	Percent
1	Nnenum	16	45	58	0	727	100.0%
2	MN BaB	16	44	57	0	715	98.3%
3	α,β-CROWN	15	45	56	1	612	84.2%
4	Verinet	16	43	0	0	590	81.2%
5	Peregrinn	14	42	0	0	560	77.0%
6	Cgdtest	1	42	43	15	-984	0%

Table 6: Benchmark `mnist-fc`

#	Tool	Verified	Falsified	Fastest	Penalty	Score	Percent
1	α,β-CROWN	66	18	53	0	963	100.0%
2	Verinet	53	18	50	0	817	84.8%
3	MN BaB	53	18	47	0	804	83.5%
4	Debona	48	18	38	0	737	76.5%
5	Nnenum	48	11	29	0	649	67.4%
6	Marabou	44	16	0	0	600	62.3%
7	Peregrinn	27	11	7	0	394	40.9%
8	Cgdtest	66	3	23	5	241	25.0%
9	Verapak	40	2	42	4	104	10.8%

Table 7: Benchmark `nn4sys`

#	Tool	Verified	Falsified	Fastest	Penalty	Score	Percent
1	α,β-CROWN	152	0	132	0	1799	100.0%
2	MN BaB	106	0	8	0	1140	63.4%
3	Verinet	57	0	43	0	661	36.7%
4	Peregrinn	24	0	22	0	284	15.8%
5	Nnenum	23	0	8	0	246	13.7%
6	Debona	2	0	2	0	24	1.3%
7	Cgdtest	2	0	2	2	-176	0%

Table 8: Benchmark `oval21`

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

Table 9: Benchmark `reach-prob-density`

#	Tool	Verified	Falsified	Fastest	Penalty	Score	Percent
1	Nnenum	22	14	22	0	411	100.0%
2	α,β-CROWN	22	14	23	0	406	98.8%
3	Verinet	22	14	10	0	383	93.2%
4	MN BaB	22	12	14	0	368	89.5%
5	Marabou	17	14	12	0	334	81.3%
6	Peregrinn	18	14	2	0	324	78.8%
7	Cgdtest	0	5	5	0	60	14.6%
8	Debona	0	2	2	0	24	5.8%

Table 10: Benchmark `rl-benchmarks`

#	Tool	Verified	Falsified	Fastest	Penalty	Score	Percent
1	α,β-CROWN	193	103	296	0	3552	100.0%
2	Verinet	193	103	292	0	3547	99.9%
3	MN BaB	193	103	288	0	3536	99.5%
4	Nnenum	191	103	283	0	3506	98.7%
5	Peregrinn	193	103	271	0	3502	98.6%
6	Marabou	191	103	278	0	3496	98.4%
7	Debona	153	99	240	0	3000	84.5%
8	Averinn	92	8	16	0	1032	29.1%
9	Cgdtest	10	24	29	3	98	2.8%
10	Verapak	0	4	0	0	40	1.1%

Table 11: Benchmark `sri-resnet-a`

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

Table 12: Benchmark `sri-resnet-b`

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

Table 13: Benchmark `tllverifybench`

#	Tool	Verified	Falsified	Fastest	Penalty	Score	Percent
1	Fastbatllnn	11	21	32	0	384	100.0%
2	MN BaB	11	21	21	0	364	94.8%
3	α,β-CROWN	11	21	12	0	353	91.9%
4	Peregrinn	10	21	7	0	324	84.4%
5	Verinet	11	21	0	0	320	83.3%
6	Nnenum	1	21	10	0	240	62.5%
7	Debona	0	19	10	0	210	54.7%
8	Marabou	4	15	2	0	194	50.5%
9	Cgdtest	0	9	6	1	2	0.5%

Table 14: Benchmark `vggnet16-2022`

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

3 Unscored Benchmarks

Table 15: Benchmark `acasxu`

#	Tool	Verified	Falsified	Fastest	Penalty	Score	Percent
1	Nnenum	139	47	174	0	2218	100.0%
2	α,β-CROWN	139	46	59	0	2021	91.1%
3	MN BaB	110	46	52	0	1664	75.0%
4	Cgdtest	85	30	115	7	680	30.7%

Table 16: Benchmark `cifar2020`

#	Tool	Verified	Falsified	Fastest	Penalty	Score	Percent
1	Verinet	91	35	109	0	1486	100.0%
2	α,β-CROWN	95	34	78	0	1479	99.5%
3	MN BaB	93	28	26	0	1275	85.8%
4	Nnenum	66	19	0	0	850	57.2%
5	Cgdtest	63	26	5	6	305	20.5%
6	Verapak	0	15	1	0	152	10.2%
7	Marabou	4	0	0	1	-60	0%

4 Stats

Table 17: Overhead

#	Tool	Seconds
1	Marabou	0.2
2	Fastbatllnn	0.5
3	Nnenum	0.9
4	Cgdtest	1.3
5	Peregrinn	1.3
6	Debona	2.0
7	Averinn	3.1
8	Verinet	3.4
9	Verapak	4.6
10	α,β-CROWN	6.7
11	MN BaB	8.2

Table 18: Num Benchmarks Participated

#	Tool	Count
1	Verinet	13
2	MN BaB	13
3	α,β-CROWN	13
4	Cgdtest	12
5	Nnenum	9
6	Marabou	8
7	Peregrinn	7
8	Debona	5
9	Verapak	3
10	Fastbatllnn	1
11	Averinn	1

Table 19: Num Instances Verified

#	Tool	Count
1	α,β-CROWN	950
2	MN BaB	812
3	Verinet	754
4	Nnenum	515
5	Marabou	514
6	Peregrinn	478
7	Cgdtest	405
8	Debona	341
9	Verapak	117
10	Averinn	100
11	Fastbatllnn	32

Table 20: Num SAT

#	Tool	Count
1	Verinet	228
2	MN BaB	227
3	α,β-CROWN	227
4	Nnenum	196
5	Peregrinn	191
6	Marabou	148
7	Debona	138
8	Cgdtest	101
9	Fastbatllnn	21
10	Averinn	8
11	Verapak	6

Table 21: Num UNSAT

#	Tool	Count
1	α,β-CROWN	723
2	MN BaB	585
3	Verinet	526
4	Marabou	366
5	Nnenum	319
6	Cgdtest	304
7	Peregrinn	287
8	Debona	203
9	Verapak	111
10	Averinn	92
11	Fastbatllnn	11

Table 22: Incorrect Results (or Missing CE)

#	**Tool**	**Count**
1	Cgdtest	41
2	Marabou	12
3	Verapak	5
4	α,β-CROWN	1