Taming Discrete Integration via the Boon of Dimensionality

Jeffrey M. Dudek, Dror Fried, Kuldeep S. Meel

Neural Network (Log-Linear) Robustness

What is the probability that an input image sampled **from a log-linear distribution** is adversarial for a given neural network?

(Baluta *et. al,* CCS 2019)

Discrete Integration

$\mathbb{P}(\varphi)$: How many **weighted** discrete solutions does a set of equations φ have?

GANAK | Dsharp + D2C | WISH

✗ Does not scale

Our Contribution: DeWeight

Add new variables/clauses to **exactly** simulate weights

Example: Discrete Integration Query

$$\varphi = (x_1 \vee x_2)$$

$$\mathbb{P}(x_1) = 2/5 \qquad \mathbb{P}(x_2) = 1/3$$

For each variable x_i:

1. Build formulas A_i and A_i' over same, new variables with $\#A_i/(\#A_i + \#A_i') = \mathbb{P}(x_i)$
2. Add clauses $(x_i \rightarrow A_i)$ and $(\neg x_i \rightarrow A_i')$

Example: Unweighted Model Counting Query

$$\varphi' = \varphi \wedge (x_1 \rightarrow A_1) \wedge (\neg x_1 \rightarrow A_1') \wedge \cdots$$

$$A_1 = y_1 \qquad A_2 = y_3$$

$$A_1' = y_1 \vee y_2 \qquad A_2' = T$$

$$\mathbb{P}(\varphi) = \frac{\#\varphi'}{\prod_i (\#A_i + \#A_i')} = \frac{9}{(2+3) * (1+2)} = \frac{9}{15}$$

Neural Network (Uniform) Robustness

What is the probability that an input image sampled **uniformly from all inputs** is adversarial for a given neural network?

(Baluta *et. al,* CCS 2019)

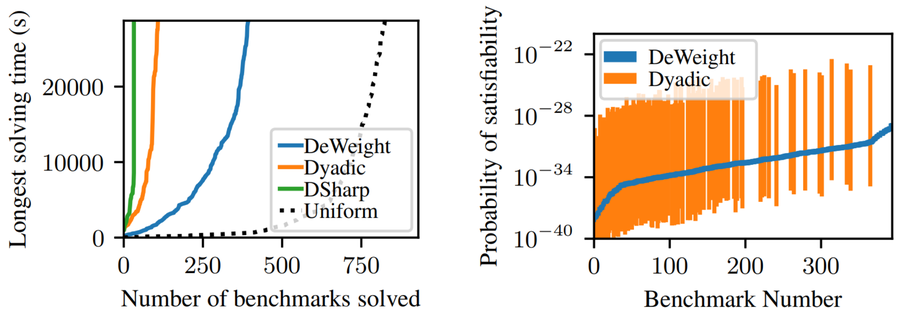
Unweighted Projected Model Counting

$\#\varphi'$: How many **unweighted** discrete solutions does a set of equations φ' have?

ApproxMC4

✓ Scales

Experimental Results with 1-decimal weights: for each pixel x of the input image, $\mathbb{P}(x \text{ is on}) \in \{0.1, \ldots, 0.9\}$



DeWeight solves **more benchmarks** (left) with **tighter formal guarantees** (right) than competing approaches.

References

[Baluta *et. al*] Teodora Baluta, Shiqi Shen, Shweta Shinde, Kuldeep S. Meel, Prateek Saxena (2019). "Quantitative Verification of Neural Networks And its Security Applications." In Proc. of CCS.

[ApproxMC4] Mate Soos and Kuldeep S Meel (2019). "BIRD: Engineering an efficient CNF-XOR SAT solver and its applications to approximate model counting." In: Proc. of AAAI.

[Dyadic] Supratik Chakraborty, Dror Fried, Kuldeep S Meel, and Moshe Y Vardi (2015). "From weighted to unweighted model counting." In: Proc. of IJCAI.

[GANAK] Shubham Sharma, Subhajit Roy, Mate Soos, and Kuldeep S. Meel (2019). "GANAK: A scalable probabilistic exact model counter." In: Proc. of IJCAI.

[Dsharp + D2C] Rehan Abdul Aziz, Geoffrey Chu, Christian Muise, and Peter Stuckey (2015). "#∃SAT: Projected model counting." In: Proc. of SAT.

[WISH] Stefano Ermon, Carla P. Gomes, Ashish Sabharwal, and Bart Selman (2013). "Taming the curse of dimensionality: Discrete integration by hashing and optimization." In: Proc. of ICML.