Attachment A

Trusted AI Challenge Series
White Paper Template

Lead Principal Investigator (PI): Professor Thomas C. Henderson

Contact Information of Lead PI: *tch@cs.utah.edu, 801-581-3601*

University/Business Name and Location: Univ. of Utah, School of Computing, SLC, UT 84112

Problem: *Topic 1*

Enter Topic Title: Verification of Autonomous Systems

Project Summary/Abstract Response to Problem:

We propose to: (1- **technical objective**) solve verification problems and provide an associated uncertainty with the result, (2 - **technical approach**) use a Reluplex-like method in which the DPLL(T) component is replaced by our recently developed Probabilistic Sentence SAT (PSSAT) method, (3 – **anticipated outcome**) find undesirable autonomous system behaviors and provide statistical guarantees (using the PSSAT probabilities), and (4 – **potential impact**) provide deeper insight into the correctness and robustness of autonomous systems. [Note that the proposed approach allows solving LTL formulas.]

Project Narrative:

Intro: Following Liu et al. [1], a feed forward NN with n layers is a function $f: \Re^{k_0} \rightarrow \Re^{k_n}$, and each layer, i, is $f_i: \Re^{k_{i-1}} \rightarrow \Re^{k_i}$, and $z_i$is the hidden variable in layer i. Given a weight matrix, $W_i \in \Re^{k_i \times k_{i-1}}$, bias vector $b_i \in \Re^{k_i}$, and activation function $\sigma_i: \Re^{k_i} \rightarrow \Re^{k_i}$, then: $z_i = f_i(z_{i-1}) = \sigma_i(W_i z_{i-1} + b_i)$. An input constraint defines a set in the domain of f, $\chi \subseteq D_X$, while an output constraint is defined as a set in the image of f, $\Upsilon \subseteq D_Y$. The *verification problem* is to show that $x \in \chi \Rightarrow y = f(x) \in \Upsilon$. This formulation allows a variety of results: counterexample, adversarial and reachability. Liu et al. [1] characterize a set of methods in terms of soundness and completeness; the most related to our proposed work are Planet [3] and Reluplex [4]. Planet searches in the function space using a SAT solver, while Reluplex is based on the simplex method and provides an efficient SMT solver for the verification problem. Reluplex is exponential cost in the worst case and uses heuristics to reduce the search space.

Research Objective: Currently, Reluplex when given a theory, T (i.e., a pair $(\Sigma, \mathbf{I})$), where Σ is a signature and **I** is a class of Σ-interpretations (models), then a formula (unquantified), φ, is T-satisfiable if it is satisfied by some interpretation in **I**. We propose to use Conjunctive Normal Form formulas and allow an associated probability with each disjunction. Then, we will apply our recently developed PSSAT method [5] to determine probabilistic satisfiability of the conditions. This will allow the ultimate objective to be achieved: given a query, the probability of the query will be determined. This reduces to SAT solving if the disjunction probabilities are set to 1. The ancillary objective will be to apply this method to autonomous UAS agent behaviors for verification analysis.

Approach: Based on work funded by AFOSR ("DDDAS-based Geospatial Intelligence: UAV Flight Path Planning in Urban Environments," Dr. Frederica Darema, Program Officer), we developed the Probabilistic Sentence Satisfiability (PSSAT) approach [5] which determines consistent assignments of probabilities to logical sentences and applied it to geospatial intelligence and UAS mission problems [6,7,8,9,10,11,12]. We propose to express a verification problem in Conjunctive Normal Form (CNF) using probabilities assigned to the conjuncts to define the input and output constraints; these conjunct probabilities characterize the uncertainty of the map, f_i, for the layer. Given a CNF formula, $S = C_1 \wedge C_2 \wedge \ldots \wedge C_m$, then the probability of a query sentence can be determined by creating a system of equations: $P(C_i) = P(L_1) + P(R) - P(L_1 \wedge R)$ (linear) or $P(C_i) = P(L_1) + P(R) -$

$P(L_1)P(R)$ (nonlinear), where $C_i = L_1 \vee ... \vee L_{k_i}$, $R = L_2 \vee ... \vee L_{k_i}$ and L_j is a literal. We find a consistent probability distribution over the logical variables (atoms) when they are assumed independent, or otherwise over the complete conjunction set of the atoms. For each sentence in the verification problem, we produce an equation in terms of atoms and conditional probabilities that arises from the standard axioms of probability. This system of equations is then solved numerically (using either linear or nonlinear solvers) to get a solution consistent with the sentence probabilities. Our findings to date indicate that for independent logical variables: (a) atom probabilities which solve PSSAT also provide a PSAT solution, (b) numerical experiments demonstrate a q-superlinear convergence rate for most test cases, (c) problems with up to 1,000 variables and 300 sentences are solved. For general variable sets (i.e., variables not independent): (a) both atom and a subset of conditional probabilities must be found, (b) a solution to PSSAT does not guarantee a solution to PSAT, but most empirical results provide such a solution, (c) convergence rates for equations with non-independent variables also appear q-superlinear.

Demonstration Experiments: We propose to try these methods on a variety of problems; e.g., the MNIST handwritten database [13], autonomous UAS tactical deconfliction [14], and a high-level autonomous agent problem like Wumpus World [15,16].

MNIST: We propose to verify properties like those suggested in [17] using the framework proposed in [2]. That is, the experiments involve a set of 150 properties resulting from combinations of the networks, perturbations, and number of images.

UAS Tactical Deconfliction: We have recently developed a lane-based UAS traffic management approach [18,19,20,21,22,23,24] which reserves time in the lanes as an efficient scheduling method. However, in case a contingency arises (e.g., a UAS goes slower than planned), then we propose the Closest Point of Approach Tactical Deconfliction Method for each UAS wherein it uses sensor data to regulate its speed, and exploits the lane structure to make efficient and effective decisions. We will encode this as a neural net and use a set of verification properties as was done in [2] for an ACAS network.

Wumpus World Decision Making: This is a 4x4 grid with an agent trying to locate gold and escape without falling into a pit or meeting a Wumpus. There are 5 percepts (binary) provided to the agent and a set of actions. At each step, the agent finds itself in a grid cell with a set of percepts (breeze, stench, glitter, bump, scream), and must choose an action (Rotate 90 degrees left, Rotate 90 degree right, Move Straight Ahead, Grab, Shoot an Arrow, or Climb); reinforcement learning will be applied to find optimal policies. Actions may not have an effect on the state of the world (e.g., Grab with no gold present). The main measure of success is a score assigned based on number of actions (negative), getting the gold (large positive), getting out (positive), or dying (large negative).

For all these domains, the main performance measures will be (1) convergence rate, and (2) time taken per algorithm (with a max cutoff time). In addition, we propose to get at the issue of neuron coverage by tracking fluctuations in σ_i values; e.g., large changes in value may signal a larger contribution to the computation. The latter two experiments provide a richer field of study for complex behavioral properties, as well as LTL rules (e.g., in Wumpus World: "Do not shoot an arrow until the Wumpus location is determined"). In addition, it may be possible to relate longer sequences of actions to the selection of undesirable actions, or even *undesirable behaviors* (a specific sequence of primitive actions). We have worked in the area of behavior analysis [25,26,27,28,29,30,31,32,33], and will examine verification problems for primitive behaviors (defined as Basic Behavior Units (BBU), like a small movement) and for complex behavior (defined as sequences of BBUs, like going in circles).

Attachment A

Trusted AI Challenge Series
White Paper Template

References:

[1] Changliu Liu, Tomer Arnon, Christopher Lazarus, Christopher Strong, Clark Barrett, and Mykel J. Kochenderfer "Algorithms for Verifying Deep Neural Networks," arXiv:1903.06758, 15 October 2020.
[2] G. Katz, C. Barrett, D. Dill, K. Julian, and Mykel Kochenderfer, "Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks," arXiv:1702.01135v2 [cs.AI] 19 May 2017.
[3] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks," Advances in Neural Information Processing Systems, pages 1097-1105, 2012.
[4] D. Silver, A. Huang, C. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, and S. Dieleman, "Mastering the Game of Go with Deep Neural Networks and Tree Search" Nature, 529(7587):484-489, 2016.
[5] Thomas C. Henderson, Robert Simmons, Bernard Serbinowski, Michael Cline, David Sacharny, Xiuyi Fan and Amar Mitiche, ``Probabilistic Sentence Satisfiability: An Approach to PSAT," Artificial Intelligence, Vol. 278, January, 2020.
[6] Thomas C. Henderson, Amar Mitiche, Robert Simmons and Xiuyi Fan, ``A Preliminary Study of Probabilistic Argumentation," Technical Report UUCS-17-001, University of Utah, Salt Lake City, UT, February, 2017.
[7] David Sacharny, Thomas C. Henderson, Amar Mitiche, Robert Simmons, Taylor Welker and Xiuyi Fan, ``BRECCIA: A Multi-Agent Data Fusion and Decision Support System for Dynamic Mission Planning," 2nd Conference on Dynamic Data Driven Application Systems (DDDAS 2017), Cambridge, MA, 7-9 August, 2017.
[8] David Sacharny, Thomas C. Henderson, Amar Mitiche, Robert Simmons, Taylor Welker and Xiuyi Fan, ``BRECCIA: Unified Probabilistic Dynamic Geospatial Intelligence," IEEE Conference on Intelligent Robots and Systems (IROS 2017 Late Breaking Paper), Vancouver, Canada, 24-28 September, 2017.
[9] Thomas C. Henderson, Robert Simmons, Amar Mitiche, Xuiyi Fan and David Sacharny, ``A Probabilistic Logic for Multi-source Heterogeneous Information Fusion," IEEE Conference on Multisensor Fusion and Integration, Daegu, South Korea, 15-18 November, 2017.
[10] David Sacharny, Thomas C. Henderson, Robert Simmons, Amar Mitiche, Taylor Welker and Xiuyi Fan, ``BRECCIA: A Novel Multi-source Fusion Framework for Dynamic Geospatial Data Analysis," IEEE Conference on Multisensor Fusion and Integration, Daegu, South Korea, 15-18 November, 2017.
[11] Thomas C. Henderson, Robert Simmons, Bernard Serbinowski, Xiuyi Fan, Amar Mitiche, and Michael Cline ,``Probabilistic Logic for Intelligent Systems," International Conference on Intelligent Autonomous Systems, Baden-Baden, Germany, 11-15 June, 2018.
[12] Thomas C. Henderson and Michael Cline, ``Using NILS to Solve Probabilistic Satisfiability for CNF Knowledge Bases," University of Utah Technical Report, UUCS-18-006, December 2018.
[13] Y. LeCun and C. Cortes, "MNIST handwritten digit database," 2010.
[14] David Sacharny, Thomas C. Henderson and Ejay Guo, "A DDDAS Protocol for Real-Time UAS Flight Coordination," InfoSymbiotics/Dynamic Data Driven Applications Systems Conference, Boston, MA, October 2-4, 2020.
[15] Stuart Russell and Peter Norvig, Artificial Intelligence: A Modern Approach, Pearson Education, NJ 2021.
[16] Shar Whisenhunt and Dianne Cook, "Comparison of Techniques to Learn Agent Strategies in Adversarial Games," AAAI Technical Report WS-98-16, 1998.
[17] C. Liu and T. Johnson, "Neural Network Verifications Workshop, VNN-COMP," *International Conference on Computer-Aided Verification*, 2020.

[18] David Sacharny and Thomas C. Henderson, ``A Lane-based Approach for Large-scale Strategic Conflict Management for UAS Service Suppliers," IEEE International Conference on Unmanned Aerial Systems, Atlanta, GA, June 2019.
[19] David Sacharny, Thomas C. Henderson and Michael Cline, ``An Efficient Strategic Deconfliction Algorithm for Large-Scale UAS Traffic Management," University of Utah Technical Report, UUCS-20-010, June, 2020.
[20] David Sacharny, Thomas C. Henderson and Michael Cline, ``Large-Scale UAS Traffic Management (UTM) Structure," IEEE Multisensor Fusion and Integration Conference, Karlsruhe, Germany, September, 2020.
[21] David Sacharny, Thomas C. Henderson, Michael Cline, Benjamin Russon and Ejay Guo, ``FAA-NASA vs. Lane-Based Strategic Deconfliction," IEEE Multisensor Fusion and Integration Conference, Karlsruhe, Germany, September, 2020.
[22] David Sacharny, Thomas C. Henderson, Michael Cline, and Benjamin Russon, ``Reinforcement Learning at the Cognitive Level in a Belief, Desire, Intention UAS Agent," SoC Technical Report, UUCS-20-013, October, 2020.
[23] David Sacharny, Thomas C. Henderson, Michael Cline, and Benjamin Russon, "Reinforcement Learning at the Cognitive Level in a Belief, Desire, Intention UAS Agent," David Sacharny, Thomas C. Henderson, Michael Cline, and Benjamin Russon, Intelligent Autonomous Systems Conference, Singapore, June, 2021.
[24] David Sacharny, Thomas C. Henderson, and Vista Marston, "On-Demand Virtual Highways for Dense UAS Operations," University of Utah, Technical Report, UUCS-21-012, May, 2021.
[25] Xinwei Xue and Thomas C. Henderson, ``Feature Fusion for Basic Behavior Unit Segmentation from Video Sequences," Robotics and Autonomous Systems, Vol. 57, No. 3, pp. 239-248, March, 2009.
[26] Thomas C. Henderson and Xinwei Xue, ``Constructing Comprehensive Behavior Models," 18th International Conference on Computer Applications in Industry and Engineering (CAINE-2005), Honolulu, Hawaii, November, 2005.
[27] Xinwei Xue and Thomas C. Henderson, ``Video Based Animal Behavior Analysis," University of Utah, Technical Report, UUCS-06-006, Salt Lake City, UT, June, 2006.
[28] Xinwei Xue and Thomas C. Henderson, ``Exploration of The Vector Fusion Method for Basic Behavior Unit Segmentation from Visual Data," IEEE Conference on Multisensor Fusion and Integration, Heidelberg, Germany, September, 2006.
[29] Xinwei Xue and Thomas C. Henderson, ``Video-based Animal Behavior Analysis From Multiple Cameras," IEEE Conference on Multisensor Fusion and Integration, Heidelberg, Germany, September, 2006.
[30] Thomas C. Henderson, Mark Bradakis and Joe Zachary, ``Reactive Behavior Design Tools," IEEE Symposium on Intelligent Control, Glasgow, Scotland, UK, pp. 178-183, August, 1992.
[31] Thomas C. Henderson, Patrick Dalton and Joe Zachary, ``A Research Program for Autonomous Agent Behavior Specification and Analysis," IEEE International Symposium on Intelligent Control, Washington D.C., August, 1991.
[32] Thomas C. Henderson and Rod Grupen, ``Logical Behaviors," Journal of Robotic Systems, Special Issue on Multisensor Integration, Vol. 7, No. 3, pp. 309-336, June, 1990.
[33] Rod Grupen and Thomas C. Henderson, ``Autochthonous Behaviors: Mapping Perception to Action," in Traditional and Non-Traditional Robotic Sensors, Thomas C. Henderson (ed), Springer-Verlag, NATO ASI Series, Berlin, pp. 285-312, 1990.