Affidabilità e sicurezza dell'AI in sistemi ad elevata criticità

Giorgio Buttazzo, Tommaso Cucinotta, Alessandro Biondi, Marco Vannucci, Daniel Casini, Valentina Colla, Mauro Marinoni, Giorgiomaria Cicero, Federico Nesti, Giulio Rossolini, Fabio Brau, Gianluca D'Amico, Edoardo Cittadini
Dipartimento di Eccellenza in Robotics and AI
Scuola Superiore Sant'Anna, Pisa
Email: nome.cognome@santannapisa.it

Abstract

Le eccellenti prestazioni delle reti neurali e degli algoritmi di apprendimento automatico stanno spingendo l'industria ad adottare tale tecnologia in diversi domini applicativi, compresi quelli ad elevata criticità, come la guida autonoma, la robotica avanzata e la diagnosi automatica di immagini mediche. Tuttavia, le metodologie di intelligenza artificiale oggi disponibili non sono state progettate per funzionare in ambienti critici e presentano numerosi problemi che devono essere affrontati a diversi livelli architetturali. Questo documento descrive le ricerche in corso presso la Scuola Superiore Sant'Anna finalizzate ad aumentare la sicurezza, l'affidabilità e la prevedibilità temporale dei sistemi basati su AI.

1 Unità Lab AIIS

Scuola Superiore Sant'Anna

2 Titolo del contributo

Affidabilità e sicurezza dell'AI in sistemi ad elevata criticità

3 Persone coinvolte

1. Giorgio Buttazzo (Prof. Ordinario)
2. Tommaso Cucinotta (Prof. Associato)
3. Alessandro Biondi (RTD_B)
4. Marco Vannucci (RTD_A)
5. Daniel Casini (RTD_A)
6. Valentina Colla (Tecnico EP)
7. Mauro Marinoni (Tecnico D)
8. Giorgiomaria Cicero (Assegnista)
9. Federico Nesti (Dottorando)
10. Giulio Rossolini (Dottorando)
11. Fabio Brau (Dottorando)
12. Gianluca D'Amico (Dottorando)
13. Edoardo Cittadini (Dottorando)
14. Giacomo Lanciano (Dottorando)
15. Fabio Galli (Dottorando)

4 Temi di ricerca

L'unità ha esperienza su metodi e algoritmi per aumentare la sicurezza e l'affidabilità di reti neurali in sistemi safety-critical. In particolare, le ricerche attive riguardano:

- **Metodi di difesa per attacchi di tipo avversario**. Tale ricerca ha lo scopo di sviluppare metodi per la rilevazione di ingressi avversari, inclusi gli attacchi robusti alle trasformazioni (quali traslazione, rotazione, sfocatura, ecc.). Il metodo si basa sull'ottimizzazione di una speciale perturbazione difensiva realizzata per essere applicata all'immagine di ingresso e progettata per rimuovere la robustezza degli attacchi avversari. La ricerca esplora anche la capacità di rilevare attacchi trasferibili a diversi modelli di rete neurale, attraverso l'introduzione di reti neurali ridondanti seguite da un meccanismo di voto a maggioranza [3].

- **Analisi di copertura per aumentare l'affidabilità delle reti neurali**. La ricerca mira a rilevare input insicuri mediante metodi di analisi di copertura al momento dell'inferenza. In una fase off-line, vengono memorizzati, per ogni layer e classe, i range di attivazione dei neuroni su un dataset attendibile. Durante l'inferenza, le attivazioni relative ad un nuovo input vengono confrontate con quelle memorizzate in precedenza, quantificando la deviazione dal comportamento attendibile. Tale deviazione viene utilizzata per calcolare il livello di affidabilità della previsione. Diversi metodi di analisi della copertura sono stati valutati e testati nell'architettura per valutare logiche di rilevamento multiple [4].

- **Garanzia di robustezza contro attacchi avversari**. Derivazione teorica di un bound per valutare la robustezza di ingressi ad attacchi avversari. In particolare, la ricerca è finalizzata a stimare la distanza di un ingresso dal confine di classificazione di una rete neurale, in modo da garantire, in fase di inferenza, che l'ingresso sia genuino, assumendo attacchi avversari di entità minore della distanza stimata.

- **Attacchi e metodi di difesa per reti di semantic segmentation**. La ricerca ha lo scopo di rilevare esempi avversari nel mondo reale ottenuti mediante opportuni cartelli che possono essere realizzati per ingannare le reti di segmentazione semantica in applicazioni di guida autonoma. La ricerca ha anche lo scopo di valutare la robustezza dei modelli esistenti per segmentazione semantica, identificando i punti di debolezza [1].
- **Sicurezza in sistemi cyber-fisici basati su AI**. La ricerca mira a realizzare un'architettura software basata su hypervisor che consenta di integrare un dominio di calcolo ad alte prestazioni, che ospita algoritmi di machine learning, con un dominio di calcolo sicuro e certificabile, che ospita componenti software ad elevata criticità, come controllori, algoritmi di voting, sistemi di rilevazione di ingressi avversari o inaffidabili elaborati da reti neurali [2, 6, 12].
- **Accelerazione predicibile di reti neurali su FPGA**. La tecnologia FPGA è in grado di accelerare l'inferenza di reti neurali con una maggiore predicibilità un minore consumo energetico rispetto alle GPU. Tuttavia, uno dei principali problemi dei sistemi basati su FPGA è una maggiore complessità di progettazione del software, che richiede una notevole esperienza per gestire correttamente le risorse disponibili. Questa ricerca sfrutta la riconfigurazione parziale dinamica per creare una FPGA virtuale più grande, in cui diverse reti neurali possono essere eseguite in timesharing sul tessuto fisico dell'FPGA, fornendo un framework automatizzato per sintetizzare acceleratori neurali ottimizzati per garantire vincoli temporali e di risorse [6, 7, 9, 10, 11, 27].
- **Miglioramento della prevedibilità nei motori di inferenza**. Lo scheduler nativo utilizzato in Tensor Flow per eseguire reti neurali profonde su piattaforme multicore è ottimizzato per il caso medio, ma può introdurre ritardi imprevedibili, rendendolo inadatto all'uso per applicazioni real-time ad elevata criticità. Questa ricerca mira a modellare lo scheduler TensorFlow e migliorarne la prevedibilità, introducendo nuovi meccanismi a livello di nodo progettati per gestire il carico computazionale specifico di una rete neurale [8, 13, 14].
- **Operazioni basate sull'intelligenza artificiale dei servizi cloud**. Questa ricerca esamina e fornisce prototipi di strumenti basati sull'intelligenza artificiale per automatizzare le operazioni nei servizi cloud, inclusi il ridimensionamento elastico e la gestione dei guasti, basati sull'apprendimento continuo dei modelli di intelligenza artificiale basati sulle metriche KPI disponibili, nonché sull'apprendimento dalle azioni degli operatori [23].
- **Tecniche di AI a supporto delle operazioni di cloud computing e infrastrutture di rete virtualizzata (VNF)**. Questa ricerca indaga tecniche di machine learning e reti neurali per analizzare l'enorme quantità di dati provenienti dal sistema di monitoraggio di un'infrastruttura cloud, per scopi legati alle operazioni di supporto, risoluzione dei problemi delle prestazioni, analisi delle cause principali, previsione del carico di lavoro e pianificazione della capacità [15, 16, 17, 18, 19, 20, 21].
- **Strumenti open source per un'elaborazione SOM efficiente su piattaforme multicore e con accelerazione GPU**. In questa ricerca, è stato realizzato XPySom, un'implementazione open-source delle note mappe auto-organizzanti (SOM) progettata per ottenere prestazioni elevate su un singolo nodo, sfruttando librerie Python ampiamente disponibili per l'elaborazione tensoriale su CPU multi-core e GPGPU. La sperimentazione con il set di dati aperti Extended MNIST mostra un'accelerazione fino a 7x e 100x rispetto alle migliori implementazioni multi-core open source che si possono trovare rispettivamente con accelerazione multi-core e GPGPU, ottenendo lo stesso livello di accuratezza in termini di errore di quantizzazione [22].
- **Apprendimento federato attento alla privacy**. Questa ricerca indaga sull'applicabilità dei metodi per la privacy differenziale al contesto della formazione di modelli di IA ad apprendimento federato, in modo da preservare in modo dimostrabile la privacy dei dati degli utenti per i singoli nodi di una rete distribuita.

5 Progetti

- Predictable, Safe, and Secure Software Systems for Autonomous Driving. Industrial project funded by Huawei, 2019-2021.
- Enhanced Train localization by sensor fusion and machine learning. Industrial project funded by Hitachi Rail STS. 2020-2022.
- Temporal isolation and security for future automotive systems using virtualization and machine learning technologies. Industrial project by Marelli, 2019-2020.
- Intelligent flying vehicle for precision agriculture. Industrial project sponsored by a technology corporation, 2020-2022.
- Artificial Intelligence and Machine Learning for automated operations in data-centers for Cloud Computing and Network Function Virtualization. Industrial project by Vodafone, 2020-2022.

6 Applicazioni in sviluppo / sviluppate / operative

- Controllo neurale di un pendolo inverso in grado di commutare automaticamente verso un controllore classico nel caso in cui la rete neurale porti il sistema in una condizione a rischio di instabilità (operativa).
- Tracking visivo di oggetti in movimento con algoritmi neurali robusti in grado di tollerare forti variazioni di forma dell'oggetto inseguito (sviluppata).
- Sistema di virtualizzazione basato su hypervisor in grado di creare due domini applicativi: uno gestito dal sistema operativo Linux, per l'inferenza di reti neurali, e l'altro gestito da un sistema operativo real time, per l'esecuzione di software critico. L'hypervisor utilizza tecniche di isolmento temporale e Trusted Execution Environemnt per garantire una maggiore sicurezza in applicazioni di tipo automotive (operativa).
- Algoritmi di rilevazione di attacchi avversari nel mondo reale per semantic segmentation network per autonomous driving (sviluppata).
- Accelerazione hardware reti neurali per la classificazione di oggetti su FPGA con tempi di risposta altamente predicibili e a basso consumo energetico, da utilizzare a bordo di droni autonomi (in sviluppo).
- Rilevazione di anomalie nel traffico di rete in datacenters for Cloud Computing and Network Function Virtualization (operativa).
- Applicazioni di machine learning per la modellazione e l'ottimizzazione di processi industriali (sviluppate).

7 Challenges / prospettive

Le ricerche in corso presso il nodo potranno consentire di:

- Aumentare l'affidabilità di reti neurali in sistemi di controllo ad elevata criticità, rilevando possibili predizioni errate mediante il monitoraggio delle attivazioni neurali e commutando temporaneamente il controllo verso sistemi di backup meno sofisticati, ma più sicuri.
- Rilevare e gestire attacchi di tipo avversario verso reti neurali, rendendo più sicuri i sistemi di percezione basati su deep learning.
- Rilevare input anomali, notevolmente diversi da quelli utilizzati durante l'addestramento di una rete neurale, che potrebbero causare una risposta inaffidabile.
- Sviluppare un framework affidabile e predicibile per l'accelerazione hardware di reti neurali su architetture di calcolo eterogenee basate su FPGA, adatto per sistemi embedded alimentati a batteria, con forti vincoli di peso e consumo energetico (droni, rover, satelliti, sonde per l'esplorazione spaziale).
- Utilizzare modelli di reinforcement learning in apprendimento continuo su sistemi fisici (oltre che simulati), al fine di affinare le prestazioni del modello sulla reale dinamica del sistema.

8 Riferimenti a paper e risorse su web

1. Federico Nesti, Giulio Rossolini, Saasha Nair, Alessandro Biondi, and Giorgio Buttazzo, "Evaluating the Robustness of Semantic Segmentation for Autonomous Driving against Real-World Adversarial Patch Attacks", Proc. of the Winter Conf. on Applications of Computer Vision (WACV), Waikoloa, Hawaii, Jan 4-8, 2022.
2. Luca Belluardo, Andrea Stevanato, Daniel Casini, Giorgiomaria Cicero, Alessandro Biondi, and Giorgio Buttazzo, "A multi-domain software architecture for safe and secure autonomous driving", Proc. of the 27th IEEE International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA 2021), Online event, August 18-20, 2021.
3. Federico Nesti, Alessandro Biondi, and Giorgio Buttazzo, "Detecting Adversarial Examples by Input Transformations, Defense Perturbations, and Voting", IEEE Transactions on Neural Networks and Learning Systems, to appear. Also in arXiv:2101.11466 [cs.CV], January 2021.
4. Giulio Rossolini, Alessandro Biondi, and Giorgio Buttazzo, "Increasing the Confidence of Deep Neural Networks by Coverage Analysis", arXiv:2101.12100 [cs.LG], January 2021.
5. Marco Pacini, Federico Nesti, Alessandro Biondi and Giorgio Buttazzo, "X-BaD: A Flexible Tool for Explanation-Based Bias Detection", Proc. of the IEEE International Conference on Cyber Security and Resilience (Virtual), July 26-28, 2021.
6. Alessandro Biondi, Daniel Casini, Giorgiomaria Cicero, Niccolò Borgioli, Giorgio Buttazzo, et al., "SPHERE: A Multi-SoC Architecture for Next-generation Cyber-Physical Systems Based on Heterogeneous Platforms", IEEE Access, Vol. 9, pp. 75446-75459, May 2021.
7. Biruk Seyoum, Marco Pagani, Alessandro Biondi, Sara Balleri, and Giorgio Buttazzo, "Spatio-Temporal Optimization of Deep Neural Networks for Reconfigurable FPGA SoCs", IEEE Transactions on Computers, to appear.
8. Daniel Casini, Alessandro Biondi, and Giorgio Buttazzo, "Timing Isolation and Improved Scheduling of Deep Neural Networks for Real-Time Systems", Software: Practice and Experience, Vol. 50, Issue 9, pp. 1760-1777, September 2020.
9. Francesco Restuccia, Marco Pagani, Alessandro Biondi, Mauro Marinoni, and Giorgio Buttazzo, "Modeling and Analysis of Bus Contention for Hardware Accelerators in FPGA SoCs", Proc. of the 32nd Euromicro Conf. on Real-Time Systems, Modena, Italy, July 7-10, 2020.

10. Francesco Restuccia, Alessandro Biondi, Mauro Marinoni, and Giorgio Buttazzo, "Safely Preventing Unbounded Delays During Bus Transactions in FPGA-based SoC", Proceedings of the 28th IEEE International Symposium On Field-Programmable Custom Computing Machines (FCCM 2020), Fayetteville, Arkansas, USA, May 3-6, 2020.
11. Francesco Restuccia, Alessandro Biondi, Mauro Marinoni, Giorgiomaria Cicero, and Giorgio Buttazzo, "AXI HyperConnect: A Predictable, Hypervisor-level AXI Interconnect for Hardware Accelerators in FPGA SoC", Proceedings of the 57th ACM/IEEE Design Automation Conference (DAC 2020), San Francisco, CA, USA, July 19-23, 2020.
12. Alessandro Biondi, Federico Nesti, Giorgiomaria Cicero, Daniel Casini, and Giorgio Buttazzo, "A Safe, Secure, and Predictable Software Architecture for Deep Learning in Safety-Critical Systems", IEEE Embedded Systems Letters, Vol. 12, No. 3, September 2020.
13. Daniel Casini, Alessandro Biondi, and Giorgio Buttazzo, "Analyzing Parallel Real-Time Tasks Implemented with Thread Pools", Proc. of the 56th ACM/ESDA/IEEE Design Automation Conference (DAC 2019), Las Vegas, NV, USA, June 2-6, 2019.
14. Daniel Casini, Alessandro Biondi, and Giorgio Buttazzo, "Deep Neural Networks for Safety-Critical Applications: Vision and Open Problems", Proc. of the 9th Int. Real-Time Scheduling Open Problems Seminar (RTSOPS 2018), Barcelona, Spain, July 3, 2018 (Best paper award).
15. T. Cucinotta, G. Lanciano, A. Ritacco, F. Brau, F. Galli, V. Iannino, M. Vannucci, A. Artale, J. Barata, E. Sposato. "Forecasting Operation Metrics for Virtualized Network Functions," Proc. of the 21st IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (IEEE CCGRID 2021), May 10-13, 2021, Melbourne, Victoria, Australia.
16. G. Lanciano, A. Ritacco, F. Brau, T. Cucinotta, M. Vannucci, A. Artale, J. Barata, E. Sposato. "Using Self-Organizing Maps for the Behavioral Analysis of Virtualized Network Functions," Communications in Computer and Information Science (CCIS), Vol. 1399, 2021, Springer, Cham.
17. T. Cucinotta, M. Vannucci, G. Lanciano, F. Galli, F. Brau, A. Artale, E. Sposato, L. N. P. Jorge. "Metodo per la gestione di risorse di un'infrastruttura per la network function virtualization," Filed IT Patent 102020000031034, December 2020.
18. T. Cucinotta, G. Lanciano, A. Ritacco, M. Vannucci, A. Artale, J. Barata, E. Sposato, L. Basili. "Behavioral analysis for Virtualized Network Functions: a SOM-based approach," Proc. of the 10th Int. Conf. on Cloud Computing and Services Science (CLOSER 2020), May 7-9, 2020, Prague, Czech Republic.
19. G. Lanciano, A. Ritacco, T. Cucinotta, M. Vannucci, A. Artale, L. Basili, E. Sposato. "SOM-based behavioral analysis for virtualized network functions," Proc. of the 35th ACM/SIGAPP Symposium On Applied Computing Brno, Czech Republic, March 30th, 2020.
20. T. Cucinotta, M. Vannucci, A. Ritacco, G. Lanciano, A. Artale, J. Barata, E. Sposato. "Metodo per predire l'evoluzione temporale di una pluralità di dati relativi ad un'infrastruttura telefonica per la network function virtualization," Filed IT Patent 102019000014262, August 2019.
21. T. Cucinotta, M. Vannucci, A. Ritacco, G. Lanciano, A. Artale, J. Barata, E. Sposato. "Metodo per identificare e classificare le modalità comportamentali di una pluralità di dati relativi ad un'infrastruttura telefonica per la network function virtualization," Filed IT Patent 102019000014241, August 2019.
22. R. Mancini, A. Ritacco, G. Lanciano, T. Cucinotta. "XPySom: High-Performance Self-Organizing Maps," in Proceedings of the 32nd IEEE Int. Symposium on Computer Architecture and High Performance Computing, September 8-11, 2020. Porto, Portugal.
23. G. Lanciano, F. Galli, T. Cucinotta, D. Bacciu, A. Passarella. "Predictive Auto-scaling with OpenStack Monasca," Proc. of the 14th IEEE/ACM Int. Conf. on Cloud Computing, December 6-9, 2021, Leicester, UK.
24. Vannucci, M., Colla, V., Dettori, S., Iannino, V. Fuzzy Adaptive Genetic Algorithm for Improving the Solution of Industrial Optimization Problems, Journal of Intelligent Systems, 29 (1), pp. 409-422, 2020.
25. Vannucci, M., Colla, V., Cateni, S. Learners reliability estimated through neural networks applied to build a novel hybrid ensemble method, Neural Processing Letters, 46 (3), pp. 791-809, 2017.
26. Cateni, S., Colla, V., Vannucci, M. A method for resampling imbalanced datasets in binary classification tasks for real-world problems, Neurocomputing, 135, pp. 32-41, 2014.
27. E. Quiñones, S. Royuela, C. Scordino, L. M. Pinho, T. Cucinotta, B. Forsberg, A. Hamann, D. Ziegenbein, P. Gai, A. Biondi, L. Benini, J. Rollo, H. Saoud, R. Soulat, G. Mando, L. Rucher, L. Nogueira. "The AMPERE Project: A Model-driven development framework for highly Parallel and EneRgy-Efficient computation supporting multi-criteria optimization," Proc. of the 23rd IEEE Int. Symposium on Real-Time Distributed Computing, May 19-21, 2020, Nashville, Tennessee, USA.