



Universiteit
Leiden
The Netherlands

Automated machine learning for neural network verification

König, H.M.T.

Citation

König, H. M. T. (2025, October 9). *Automated machine learning for neural network verification*. Retrieved from <https://hdl.handle.net/1887/4266921>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/4266921>

Note: To cite this publication please use the final published version (if applicable).

Samenvatting

In een tijdperk dat gekenmerkt wordt door de wijdverspreide inzet van neurale netwerken, kan het belang van het waarborgen van hun betrouwbaarheid en veiligheid niet genoeg worden benadrukt. In dit proefschrift staat de verificatie van neurale netwerken centraal, een cruciale maar uitdagende taak gericht op het bewijzen van de robuustheid van neurale netwerkmodellen. Daarnaast introduceert deze dissertatie nieuwe technieken en strategieën die de efficiëntie van verificatiealgoritmen voor neurale netwerken aanzienlijk verbeteren en zo bijdragen aan de ontwikkeling van veiligere en betrouwbaardere toepassingen van deep learning.

Hoofdstuk 3 presenteert een gedetailleerd onderzoek van het huidige landschap in verificatie van neurale netwerken, specifiek gericht op de verificatie van lokale robuustheid. De diversiteit binnen het onderzoeksveld, zowel wat betreft verificatietechnieken als neurale netwerkarchitecturen, vormt een complexe uitdaging voor experts die de veiligheid en betrouwbaarheid van deze systemen willen garanderen. We voeren een grondige empirische analyse uit van een aantal prominente verificatiealgoritmen en brengen daarmee de gefragmenteerde aard van het onderzoeksveld aan het licht. De bevindingen suggereren dat er niet één algoritme domineert in alle verificatieproblemen. In plaats daarvan zijn de prestaties van verschillende algoritmen zeer complementair, wat wijst op de potentiële voordelen van het gebruik van algoritmeportfolio's om de verificatie-efficiëntie te verbeteren. Deze genuanceerde kijk benadrukt het belang van het overwegen van een reeks methoden en technieken om de gevarieerde uitdagingen van verschillende neurale netwerkconfiguraties en verificatie-eigenschappen aan te pakken.

Hoofdstuk 4 gaat vervolgens in op mixed integer programming (MIP)-gebaseerde verificatie van neurale netwerken, met speciale aandacht voor het versnellen van het verificatieproces door middel van geautomatiseerde algoritmeconfiguratie. Het hoofdstuk introduceert nieuwe benaderingen om algoritmeportfolio's in te zetten en laat zien hoe de prestaties van MIP-gebaseerde verificatiesystemen aanzienlijk kunnen

worden verbeterd. Door gebruik te maken van geautomatiseerde configuratietechnieken toont het hoofdstuk de mogelijkheid om verificatiemethoden beter af te stemmen op de specifieke kenmerken van een specifiek neurale netwerk, waardoor de berekenkosten worden verlaagd en de efficiëntie van de verificatietaak wordt verbeterd.

Hoofdstuk 5 introduceert een nieuw perspectief op verificatie door de voorspelende mogelijkheden te onderzoeken met betrekking tot de rekentijd van verschillende verificatiealgoritmen voor specifieke problemen. Deze benadering is erop gericht om rekenkracht adequater onder verschillende problemen toe te wijzen, door de inspanningen te richten op verificatieproblemen waarvan de kans groter is dat ze binnen redelijke tijdbudgetten kunnen worden opgelost. De integratie van voorspelling van de benodigde rekentijd in het verificatieproces vertegenwoordigt een strategische verschuiving naar methodologieën die bewust omgaan met reken middelen, die mogelijk de efficiëntie en toepasbaarheid van verificatie van neurale netwerken in echte scenario's kunnen veranderen.

In het laatste thematische hoofdstuk, hoofdstuk 6, spitst de discussie zich toe op de kritieke taak van robuuste modelselectie binnen het domein van adversariële robuustheid. We stellen een geavanceerd race-algoritme voor dat gebruik maakt van eenvoudige maar nieuwe heuristieken om efficiënt het meest robuuste neurale netwerkmodel te bepalen uit een gegeven set. Deze methode stroomlijnt niet alleen het modelselectieproces, maar vermindert ook aanzienlijk de computationele verspilling die gepaard gaat met het evalueren van meerdere kandidaatmodellen.

Concreet hebben we aangetoond hoe geautomatiseerde machine learning, of meta-algoritmische benaderingen in het algemeen, de prestaties en praktische efficiëntie van verificatiesystemen voor neurale netwerken kunnen verbeteren en daarmee bijdragen aan een veiliger en betrouwbaarder gebruik van diepe neurale netwerken in het ontwikkelende landschap van kunstmatige intelligentietoepassingen.