Soutenance de thèse

Iryna DE ALBUQUERQUE SILVA soutiendra sa thèse de doctorat, préparée au sein de l'équipe d'accueil doctoral ISAE-ONERA MOIS et intitulée «*Implémentation certifiable et efficace de réseaux de neurones sur des systèmes embarqués temps-réel critiques*»

Le 16 juillet 2024 à 10h00, salle des thèses, ISAE-SUPAERO

devant le jury composé de

M. Thomas CARLE	Université Toulouse III – Paul Sabatier	Co-directeur de thèse
M. Albert COHEN	Google France	
Mme Liliana CUCU-GROSJEAN	INRIA Paris	Rapporteure
M. Pierre-Emmanuel HLADIK	Ecole Centrale Nantes	Rapporteur
M. Victor JEGU	Airbus	
M. Tomasz KLODA	INSA Toulouse	
Mme Claire PAGETTI	ONERA	Directrice de thèse
Mme Isabelle PUAUT	Université de Rennes 1	

Résumé : Contexte. Le monde de l'aéronautique envisage l'utilisation d'algorithmes d'apprentissage automatique, et en particulier de réseaux de neurones, pour faciliter et améliorer des tâches telles que la navigation, la maintenance prédictive et le contrôle du trafic aérien. Toutefois, leur utilisation dans des systèmes critiques soulève de nombreuses questions de sécurité et de conformité aux exigences de certification. Ces dernières visent à garantir la correction de la conception jusqu'à l'implantation de la fonction attendue. Récemment de nouvelles directives, telles que le standard ARP6983 [152], sont en cours d'écriture pour compléter les textes réglementaires existant afin de permettre l'introduction d'algorithmes d'apprentissage automatique dans des systèmes avioniques. Ces directives complètent notamment la DO-178C [67], le standard de référence pour le développement de logiciels. Cette thèse étudie l'implantation en temps réel et en toute sécurité de réseaux de neurones feed-forward sur des systèmes avioniques embarqués. En particulier, nous nous concentrons sur le modèle de réseau neuronal entrainé et vérifié hors ligne, appelé modèle d'inférence. L'objectif est de définir une approche de développement logiciel de modèles d'inférence en conformité avec exigences de l'avionique.

Approche. Nous avons défini un certain nombre d'objectifs nécessaires pour garantir l'implantation sûre et correcte d'un modèle d'inférence. Ces objectifs sont un sous-ensemble de ceux listés dans les normes de certification, mais ont été identifiés comme difficiles à atteindre pour le processus traditionnel de développement de logiciels d'apprentissage automatique et sont cohérents avec les directives en cours d'élaboration, à savoir les feuilles de route de l'EASA et la norme ARP6983. Plus précisément, les objectifs sélectionnés dans la thèse sont (i) la description précise et non ambigüe de la fonctionnalité du modèle d'inférence, (ii) la préservation de la sémantique et la traçabilité de la description du modèle d'inférence jusqu'à l'exécutable final, (iii) la prédictibilité temporelle de l'implantation, et (iv) une utilisation efficace des ressources disponibles.

Contributions. Pour implanter correctement un modèle d'inférence, il est important de définir formellement sa sémantique. Cela correspond à la première contribution de la thèse et nous avons formalisé chaque couche du modèle d'inférence sous forme de fonctions mathématiques. Ensuite, nous avons développé ACETONE, de l'anglais "Avionics C code generator for Neural Networks". ACETONE est un framework qui, à partir de la description d'un modèle d'inférence, génère automatiquement un code C sémantiquement équivalent et prédictible. Nous avons évalué le framework sur un ensemble réaliste de cas d'utilisation pour valider l'approche en regard des objectifs

de certification choisis. En outre, nous avons comparé l'approche à des outils de l'état de l'art. Une fois le framework défini et disponible sur GitHub, nous avons choisi un axe d'amélioration et d'optimisation de code pour accélérer les temps de calcul. Plus précisément nous nous sommes concentrés sur les convolutions. L'idée est de reprendre les travaux classiques qui transforment une convolution en une multiplication matricielle. Cependant, les routines d'algèbre linéaire disponibles sur étagère ne sont pas compatibles avec les exigences de l'avionique. C'est la raison pour laquelle, nous avons proposé notre propre implantation prédictive, traçable et efficace d'un algorithme de multiplication matrice-matrice (GEMM) par blocs. Nous avons montré comment configurer un tel algorithme pour une cible donnée tout en proposant une méthode formelle de calcul du nombre d'accès à la mémoire et du nombre de cache misses, ce qui ouvre la voie à une analyse statique du WCET.

Summary: Context. Aeronautics envisions the use of machine learning (ML) algorithms, and in particular neural networks, to help and improve tasks such as navigation, predictive maintenance, and air traffic control. However, their use in safety-critical products raises several questions regarding compliance with normative requirements, which aim to guarantee correctness from the design to the implementation of the intended function. Thus, in order to allow the use of ML-based systems in aeronautics, guidelines, such as the ARP6983 standard, are currently being drafted. They complete the DO-178C, the reference guidance for the implementation process of software items. This thesis studies the safe real-time implementation of feed-forward deep neural networks on embedded avionics systems. In particular, we focus on the offline trained and verified neural network model, named inference model. The purpose of this work is then to provide an approach that enables the implementation of the inference model in compliance with avionics requirements.
Approach. We define a number of quantifiable objectives necessary to guarantee the inference model's safe and correct implementation. These objectives do not represent all the requirements present in normative standards, but are recognized as complex to obtain for the traditional ML software development process and are in accordance with the guidelines under construction, i.e. EASA roadmaps and the ARP6983 standard. We notably identify certification objectives regarding (i) a precise description of the inference model's functionality, (ii) the semantics preservation and traceability between the inference model description and the final executable, (iii) the timing predictability of the implementation, and (iv) an efficient usage of the available resources.
Contributions. To correctly implement the inference model an important initial requirement is to have its semantics formally defined. Hence, our first contribution concerns the formalization of the semantics of each layer of the inference model as mathematical functions. Afterwards, we proceed to develop ACETONE, which stands for "Avionics C code generator for Neural Networks". ACETONE is a framework that, from the inference model description, automatically generates a semantically equivalent and predictable C code. We evaluate the framework on a realistic set of use cases to verify the compliance of our approach with the derived certification objectives. Moreover, we are able to compare the proposed approach against two state-of-the-art inference tools. Lastly, we work on improving the code generated by ACETONE, notably concerning the implementation of resource-expensive functions, in particular of the Convolutional layer. We focus on a Convolutional layer optimization that interprets the convolution operator as a matrix-matrix multiplication. However, state-of-the-art implementations of linear algebra routines rely on libraries that are not compatible with avionics requirements. We thus propose a predictable and traceable yet efficient implementation of a blocked general matrix-matrix multiplication (GEMM) algorithm. We then provide a set of rules for tuning the algorithm's blocking parameters and predict with precision its number of memory accesses and cache misses, which paves the way for a static WCET analysis.
Keywords: machine learning, embedded systems, certification