UNIVERSIDADE FEDERAL DO AMAZONAS
FACULDADE DE TECNOLOGIA
PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

FORMAL SYNTHESIS OF QUANTIZED NEURAL NETWORKS: AN SMT-BASED APPROACH FOR BIT-PRECISE VERIFICATION

João Victor Lima de Souza

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica, PPGEE, da Universidade Federal do Amazonas, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Elétrica.

Orientador: Lucas Cordeiro

Manaus
Março de 2025

Ministério da Educação
Universidade Federal do Amazonas
Coordenação do Programa de Pós-Graduação em Engenharia Elétrica

FOLHA DE APROVAÇÃO

Poder Executivo Ministério da Educação
Universidade Federal do Amazonas
Faculdade de Tecnologia
Programa de Pós-graduação em Engenharia Elétrica

ANTONIO MARCOS DA COSTA PEREIRA

CLASSIFICAÇÃO AUTOMÁTICA DE MODULAÇÕES EM RECEPTORES ÓPTICOS COERENTES FLEXÍVEIS

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Amazonas, como requisito parcial para obtenção do título de Mestre em Engenharia Elétrica na área de concentração Controle e Automação de Sistemas.

Aprovada em 15 de março de 2024.

Documento assinado eletronicamente por:
Prof. Dr. Waldir Sabino da Silva Júnior Presidente
Prof. Dr. Celso Barbosa Carvalho (Membro Titular 2 - Interno)
Prof. Dr. Alexandre Coutinho Matheus (Membro Titular 1 - Externo)

Manaus, 11 de junho de 2024.

Documento assinado eletronicamente por **Waldir Sabino da Silva Júnior**, **Professor do Magistério Superior**, em 11/06/2024, às 09:02, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Alexandre Coutinho Mateus**, **Usuário Externo**, em 11/06/2024, às 09:06, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Celso Barbosa Carvalho**, **Professor do Magistério Superior**, em 11/06/2024, às 15:27, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://sei.ufam.edu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2088836** e o código CRC **9C9BEB4A**.

Av. Octávio Hamilton Botelho Mourão - Bairro Coroado 1 Campus Universitário Senador Arthur Virgílio Filho, Setor Norte - Telefone: (92) 3305-1181
CEP 69080-900 Manaus/AM - mestrado_engeletrica@ufam.edu.br

Referência: Processo nº 23105.011910/2024-21 SEI nº 2088836

Criado por 62136828215, versão 3 por 62136828215 em 11/06/2024 08:47:49.

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Pereira, Antonio Marcos da Costa
P436c Classificação Automática de Modulações em Receptores Ópticos Coerentes Flexíveis / Antonio Marcos da Costa Pereira . 2024
108 f.: il. color; 31 cm.

Orientador: Waldir Sabino da Silva Júnior
Coorientador: André Luiz Aguiar da Costa
Dissertação (Mestrado em Engenharia Elétrica) - Universidade Federal do Amazonas.

1. Classificação Automática de Modulação. 2. Redes Ópticas. 3. Random Forest. 4. Perceptron Multicamada. I. Silva Júnior, Waldir Sabino da. II. Universidade Federal do Amazonas III. Título

Gostaria de expressar minha gratidão...

Agradecimentos

Expresso minha profunda gratidão...

Resumo da Dissertação apresentada à UFAM como parte dos requisitos necessários para a obtenção do grau de Mestre em Engenharia Elétrica

FORMAL SYNTHESIS OF QUANTIZED NEURAL NETWORKS: AN SMT-BASED APPROACH FOR BIT-PRECISE VERIFICATION

João Victor Lima de Souza

Orientador: Lucas Cordeiro

Programa: Pós-Graduação em Engenharia Elétrica

Palavras-chave: Classificação Automática de Modulação, Redes Ópticas, *Random Forest*, *Perceptron* Multicamada.

Abstract of Dissertation presented to UFAM as a partial fulfillment of the requirements for the degree of Master in Electrical Engineering

FORMAL SYNTHESIS OF QUANTIZED NEURAL NETWORKS: AN SMT-BASED APPROACH FOR BIT-PRECISE VERIFICATION

João Victor Lima de Souza

Advisor: Lucas Cordeiro

Department: Postgraduate in Electrical Engineering

Keywords: Automatic Modulation Classification, Optical Networks, Random Forest, Multilayer Perceptron.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introduction

1.1 Background

From the first Artificial Neural Network (ANN) models such as Perceptron, which was proposed in 1958 [Rosenblatt, 1958], to the current state-of-the-art models like GPT-4 [OpenAI, 2023], ANN has become a transformative force across various sectors. It is used in healthcare for diagnostics such as Parkison disease detection [Luo et al., 2025], in autonomous vehicles for safety features using Convolutional neural networks (CNNs) to detect driver distraction in real time [Lei et al., 2025], and in finance for fraud detection by analyzing transaction patterns using advanced machine learning algorithms [Zhu et al., 2024]. These examples illustrate the profound impact of these models on our daily lives, improving efficiency, safety, and decision-making processes.

However, these state-of-the-art ANNs are often too large and complex to be deployed on devices with limited resources, such as smartphones and IoT devices [Zhu et al., 2020]. The computational demands of these models can lead to significant latency and energy consumption, making them impractical for real-time applications in resource-constrained environments.

To address these challenges, quantization techniques are commonly used to reduce the size of models [Zhu and Gupta, 2016, Zhu et al., 2020]. It involves reducing the precision of the model's weights, activations, and bias, which can lead to a significant reduction in memory usage and computational requirements. However, this process often results in a trade-off between model size and accuracy, as lower precision can lead to degration of performance and therefore its accuracy [Zhou et al., 2017].

In works such as [Cai et al., 2025], the authors have highlighted two critical safety properties of QNNs: robustness and backdoor-freeness. The robustness would refer to the model's ability to resist small input perturbations that change the classification, while the former refers to the ability of not having backdoors, which could be intentionally explored with malicious intent. These properties are of utmost importance to ensure the reliablity and safety of QNNs.

1.2 Motivation

Thus, the need to ensure the reliability of ANN models after going through quantization is demonstrated since these models are being employed in critical applications, e.g., healthcare, autonomous vehicles, financial system fraud detection, among others [OpenAI, 2023, Luo et al., 2025, Lei et al., 2025, Zhu et al., 2024]. In these scenarios, the models must not only be efficient to run on resource-constrained environments but also maintain their robustness and safety properties while keeping it's accuracy.

Modern quantization techniques, such as those proposed by [Zhu et al., 2020, Han et al., 2015, Han et al., 2020, Jacob et al., 2018, Cai et al., 2020, Zhou et al., 2017], aim to reduce the loss of accuracy while minimizing the size of the model by employing a different set of strategies such as weight sharing, pruning, and factorization of low rank. However, these techniques often do not guarantee that the quantized model will retain its robustness or be backdoor-free.

In order to ensure that the quantization process does not compromise the model's performance, some verification techniques such as Formal Methods are employed. These methods can be used to verify the integrity of the quantized model by checking if the quantized model still meets the safety properties like the original model.

1.3 Problem Statement

Most formal verification techniques for neural networks (such as Reluplex and Marabou) assume that networks operate with real-number arithmetic [Katz et al., 2017, Amir et al., 2021]. In contrast, actual hardware implementations use finite-precision arithmetic, such as low-precision floating-point or, more frequently, fixed-point [Han et al., 2020].

Furthermore, quantization, while beneficial for efficiency, can degrade accuracy and, more critically, compromise desired safety properties. Existing works on QNN verification generally focuses on post-hoc analyses; that is, they verify a network after it has been quantized [Eleftheriadis et al., 2022, Song et al., 2023, Katz et al., 2017, Baranowski et al., 2020, Pulina and Tacchella, 2012, Cordeiro et al., 2025] aiming to ensure that the quantization process does not introduce vulnerabilities or degrade of accuracy. However, these approaches do not guarantee the optimal quantization strategy unlinke the work of [Abdi et al., 2021, Cai et al., 2025], which proposes a framework to find a optimal bit-width for each layer of a QNN.

In contrast, the use of Mixed Integer Linear Programming (MILP) solvers has been proposed to verify the preimage of QNNs [Cai et al., 2025]. This approach allows for the verification of robustness and backdoor-freeness properties of QNNs. While having achieved promising results in preserving these properties, it does not consider the fixed-point precision used in QNNs as done by [Baranowski et al., 2020].

Since the verification problem in [Cai et al., 2025] uses MILP solvers, it does not formally incorporate the nuances of fixed-point arithmetic during verification. This work aims to fill this gap by integrating the SMT Theory of Fixed-Point Arithmetic, as formalized by [Baranowski et al., 2020], into the preimage calculation and quantization process. This integration will allow the synthesis of quantization strategies that are certified for specific finite precision, ensuring that properties hold even with the effects of round and overflow.

1.4 Objectives

The goal of this work is to develop a end-to-end framework for the synthesis of quantization strategies that guarantee the preservation of robustness and backdoor-freeness properties after quantization, while directly addressing finite-precision arithmetic. This will differ form post-hoc approaches and quantization techniques that focus solely on accuracy as the primary goal will be to ensure a optimal quantization bit-width precision for each layer of the QNN. To formally achieve the proposed goal, the following specific objectives have been defined:

- **Formalize the SMT Theory to encode Quantization Synthesis:** Integrate the SMT Theory into the preimage calculation and quantization process, allowing for the synthesis of quantization strategies that are certified for specific finite precision.

- **Adapt the MILP Formulation to SMT:** Adapt the Mixed-Integer Linear Programming (MILP) formulation used in [Cai et al., 2025] to a SMT formulation that reflects fixed-point operations, ensuring that the preimage calculation accounts for the effects of finite precision.

- **Validade Framework:** Validate the proposed framework by applying it to a set of NNs, demonstrating its effectiveness in synthesing quantization that preserve the desired properties.

- **Evaluate Performance:** Evaluate the performance of the proposed framework in terms of scalability and efficiency, comparing it with existing approaches.

1.5 Contributions

From this work we expect to contribuite to the field of QNN verification by providing a framework that allows for the synthesis of quantization strategies that take into account the fixed-point precision used by QNN models. The main contributions will be:

- A formalization of the SMT Theory of Fixed-Point Arithmetic to encode quantization synthesis, allowing for the synthesis of quantization strategies that are certified for specific finite precision.

- An adaptation of the MILP formulation used in [Cai et al., 2025] to a SMT formulation as a set of contrains that reflects fixed-point operations, ensuring that the preimage calculation accounts for the effects of finite precision.

- A validation of the proposed framework by applying it to a set of NNs, demonstrating its effectiveness in synthesizing quantization strategies that preserve the desired properties.

- An evaluation of the performance of the proposed framework in terms of scalability and efficiency, comparing it with existing approaches.

1.6 Dissertation Structure

The document structure unfolds as follows. Chapter 1 introduces the background, motivations, and objectives of the work. Chapter **??** delves into the theoretical foundations, including the SMT Theory, Neural Networks, and Quantization.

Capítulo 2

Related Works

In this section we are going to review some of the most relevant works in the literature that address the optimization of QNN bit-widths while also providing formal guarantees on the network's behavior.

2.1 Certified Quantization Strategy Synthesis for Neural Networks

The work introduces Quadapter, a end-to-end framework for quantizing neural networks by choosing per-layer fixed-point precisions that provably preserve a desired property of robustnes and backdoor-freeness. Within the work, NN are written as a composition of affine and activation layers $N = f_{2d} \circ \cdots \circ f_1$. For every even affine layer $2i$, the authors compute a safe preimage $\mathcal{P}^{2i}$ such that any activation $x_{2i} \in \mathcal{P}^{2i}$ is guaranteed to map [Henzinger et al., 2020], through the suffix subnetwork $N_{[2i+1:2d]}$, into the target output set O, i.e, we garantee that for a given input interval, the output will remain in the desired output set. Preimages are built by solving a layer-wise MILP that maximizes the size of a box-like template T^{2i} under the inclusion constraint above; inclusion is checked by negating it and asking the MILP for a violating property. Given these backward preimages, the method then selects, for each affine layer, a bit-width configuration (Q_i, F_i) so that a sound forward over-approximation of the quantized activations remains inside $\mathcal{P}^{2i}$; if the MILP-based negated inclusion is unsatisfiable at every certified layer, the whole quantized network is certified for the property. The approach yields compact bit-width vectors on MNIST-like models while maintaining the specified property, and it cleanly separates (i) tight but costlier under-approximations of backward preimages MILP from (ii) fast over-approximations of forward reach sets for inclusion checks. Limitations include reliance on MILP scalability and the need to align forward abstractions with hardware semantics.

The work lacks bit-accurate guarantee for the quantized network, as the forward abstraction does not model rounding and saturation like [Baranowski et al., 2020]. Therefore, w good contribution would be to extend the method to use SMT-based abstraction to check the fixed-point constrains of rounding/saturation.

2.2 Counter Example Guided Search for Neural Network Quantization

Based on the works of [Katz et al., 2017, Katz et al., 2019] the authors propose CEG4N, an end-to-end framework for quantization of neural networks with formal guarantees of accuracy. The framework is divided into 3 main steps: (i) Bit Search Module (BSM) based on genetic algorithm that receives counter-example (H^{+1}_{CE}) from the verifier (iii); (ii) Translater that converts both NN and QNN to ONNX format to the input format of the verifier; (iii) Verifier that checks if the QNN is equivalent to the NN within a given tolerance ϵ. The verifier is based on an SMT solver (ESBMC) that encodes the equivalence property as a set of constraints $\phi := \phi \wedge \phi' \wedge \neg(y \approx_\epsilon y')$ where ϕ and ϕ' encode the NN and QNN respectively, and $y \approx_\epsilon y'$ encodes the tolerance condition. The verifier returns a counter-example H^{+1}_{CE} if the property is violated, which is then used by the BSM to generate a new bit-width configuration. The process is repeated until either a valid configuration is found or a maximum number of iterations is reached. The authors evaluate their framework on several benchmarks, including MNIST and CIFAR-10, demonstrating its effectiveness in finding low-bit configurations while maintaining accuracy within the specified tolerance.

The main limitation of the work is the scalability of the verifier, which can become a bottleneck for larger networks. For future works, the authors suggest exploring quantization approachs that operate entirelly in integer domain, pontentially improving the scalability.

Capítulo 3

Methodology

In this chapter, we intend to detail the pillar of our proposed methodology for our end-to-end framwork that aims to find the optimal bit-width for each layer of QNN while ensuring that the defined bit-widths satisfy robustness properties using SMT Theory of Fixed-Point Arithmetic.

From our work we aim to develop a end-to-end based on the work of Quadapter [Cai et al., 2025] for the synthesis of quantization and therefore find the optimal bit-width for each layer. However, we will extend it by integreting the SMT Theory of Fixed-Point Arithmetic like done in previous works [Baranowski et al., 2020, Sena et al., 2021, Zhang et al., 2020, Huang et al., 2023, Henzinger et al., 2022] to check if the defined constrains are satisfied by the quantization strategy. Whether not, the counterexample will be used to refine the quantization.

The work methodology can be described in the diagram shown in Figure 3.1. From it, we can divide the work into two sub-problems. The first problem would be to find out the minimum bit-width and it will be described in Section 3.1. The second problem would be to verify whether the defined bit-widths satisfy the required properties and it will be described in Section 3.2.

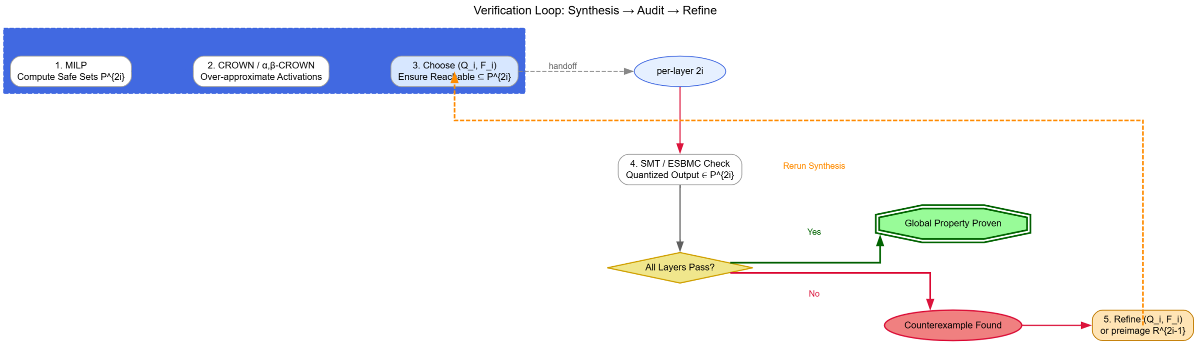

Figura 3.1: Proposed Methodology Diagram

3.1 Finding the Minimum Bit-width for Each Layer

There are several works that focus on finding the optimal bit-width for each layer of a neural network [Sena et al., 2021]. Sena et al. proposed a Bit Search Module (BSM) which uses a counter-example guided approach to find the optimal bit-width. The BSM proposes the usage of genetic algorithms that iteractively proposes a vector of bit-widths for each layer and then uses a verifier to check if the defined constrains are satisfied. If not, the counter-example is used to refine the bit-widths. The process is repeated until a satisfactory solution is found or a limit of iteractions is met.

Given a DNN N with $2d$ layers.

Our objective would be to find the minimum bit-width for each layer such that the quantized network $\hat{N}$ preserves a desired property, such as adversarial robustness or the absence of backdoors.

The vector of bit-widths for each layer if defined as $\mathbf{b} = (b_1, b_2, \ldots, b_{2d})$, where b_i is the bit-width for layer i. The quantized network $\hat{N}$ is obtained by applying a quantization function Q_{b_i} to the weights and activations of each layer i.

3.2 Verification of QNN Bit-width using Bit-precision SMT-Solving

3.2.1 Refine

Referências Bibliográficas

[Abdi et al., 2021] Abdi, M., Ravanbakhsh, M., Samudrala, P., and Chen, M. (2021). Counterexample guided neural network quantization refinement (ceg4n). In *Automated Technology for Verification and Analysis*, pages 284–300. Springer. 3

[Amir et al., 2021] Amir, G., Wu, H., Barrett, C., and Katz, G. (2021). An smt-based approach for verifying binarized neural networks. In *Tools and Algorithms for the Construction and Analysis of Systems*, pages 3–21. Springer. 2

[Baranowski et al., 2020] Baranowski, M., Bjørner, N., and de Moura, L. (2020). An smt theory of fixed-point arithmetic. In *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, pages 3–20. Springer. 3, 7, 8

[Cai et al., 2020] Cai, Y., Wang, T., Zhang, C., Chen, C., Lin, L., and Han, S. (2020). Certified Quantization: Towards Accurate and Compact Neural Networks. *arXiv preprint arXiv:2002.08153*. 2

[Cai et al., 2025] Cai, Y., Wang, T., Zhang, C., Chen, C., Lin, L., and Han, S. (2025). Certified quantization strategy synthesis for neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 36(1):1–14. 2, 3, 4, 8

[Cordeiro et al., 2025] Cordeiro, L. C., Daggitt, M. L., Girard-Satabin, J., Isac, O., Johnson, T. T., Katz, G., Komendantskaya, E., Lemesle, A., Manino, E., Šinkarovs, A., and Wu, H. (2025). Neural network verification is a programming language challenge. 3

[Eleftheriadis et al., 2022] Eleftheriadis, C., Kekatos, N., Katsaros, P., and Tripakis, S. (2022). On neural network equivalence checking using smt solvers. 3

[Han et al., 2020] Han, S., Liu, H., Mao, H., Pu, J., Pedram, M., Horowitz, M., and Dally, W. J. (2020). Understanding the impact of precision quantization on the accuracy and energy of neural networks. In *2020 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pages 468–481. IEEE. 2

[Han et al., 2015] Han, S., Mao, H., and Dally, W. J. (2015). Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*. 2

[Henzinger et al., 2020] Henzinger, T. A., Lechner, M., Noll, D., and Sezgin, C. I. (2020). On preimage approximation for neural networks. In *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, pages 188–206. Springer. 6

[Henzinger et al., 2022] Henzinger, T. A., Lechner, M., Noll, D., and Sezgin, C. I. (2022). Scalable verification of quantized neural networks (technical report). arXiv:2012.08185 [cs]. 8

[Huang et al., 2023] Huang, D. A., Ibeling, D., Katz, G., Julian, K., Lim, R., Shah, P., Thakoor, S., Wu, H., and Yad_Gar, A. (2023). Towards efficient verification of quantized neural networks. arXiv:2312.12679 [cs]. 8

[Jacob et al., 2018] Jacob, B., Kligys, S., Chen, B., Tang, M., Howard, A. G., and Kalenichenko, D. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. *arXiv preprint arXiv:1712.05877*. 2

[Katz et al., 2017] Katz, G., Barrett, C., Dill, D. L., Julian, K., and Kochenderfer, M. J. (2017). Reluplex: An efficient smt solver for verifying deep neural networks. In *Computer Aided Verification*, pages 97–117. Springer. 2, 3, 7

[Katz et al., 2019] Katz, G., Huang, D. A., Ibeling, D., Julian, K., Lazarus, C., Lim, R., Shah, P., Thakoor, S., Wu, H., Yad_Gar, A., et al. (2019). The marabou framework for verification and analysis of deep neural networks. In *Computer Aided Verification*, pages 443–452. Springer. 7

[Lei et al., 2025] Lei, J., Ni, Z., Peng, Z., Hu, H., Hong, J., Fang, X., Yi, C., Ren, C., and Wasaye, M. (2025). An intelligent network framework for driver distraction monitoring based on res-se-cnn. *Scientific Reports*, 15. 1, 2

[Luo et al., 2025] Luo, Y., Huang, Y., Fang, G., Tan, N., and Tang, Y. (2025). Efficient recognition of Parkinson's disease mice on stepping characters with CNN. *Scientific Reports*, 15(1):2566. 1, 2

[OpenAI, 2023] OpenAI (2023). GPT-4 Technical Report. Technical Report arXiv:2303.08774, OpenAI. 1, 2

[Pulina and Tacchella, 2012] Pulina, L. and Tacchella, A. (2012). Challenging smt solvers to verify neural networks. *AI COMMUNICATIONS*, 25:117–135. 3

[Rosenblatt, 1958] Rosenblatt, F. (1958). The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review*, 65(6):386–408. 1

[Sena et al., 2021] Sena, L. H., Song, X., da S Alves, E. H., Bessa, I., Manino, E., and Cordeiro, L. C. (2021). Verifying quantized neural networks using smt-based model checking. *arXiv preprint arXiv:2106.05997*. 8, 9

[Song et al., 2023] Song, X., Sun, Y., Mustafa, M. A., and Cordeiro, L. C. (2023). Qnn-repair: Quantized neural network repair. *arXiv preprint arXiv:2306.13793*. 3

[Zhang et al., 2020] Zhang, C., Chen, I.-C., and Cheng, C.-H. (2020). Qvip: An ilp-based formal verification approach for quantized neural networks. In *Computer Aided Verification*, pages 453–474. Springer. 8

[Zhou et al., 2017] Zhou, A., Yao, A., Guo, Y., Yang, Y., and Chen, Y. (2017). Incremental network quantization: Towards lossless CNNs with low-precision weights. *arXiv preprint arXiv:1702.03044*. 1, 2

[Zhu and Gupta, 2016] Zhu, Y. and Gupta, S. (2016). Trained quantization for deep neural networks. *arXiv preprint arXiv:1612.06170*. 1

[Zhu et al., 2020] Zhu, Y., Wang, X., Li, J., Zhang, Y., and Liu, H. (2020). A survey of quantization methods for efficient neural network inference. *IEEE Transactions on Neural Networks and Learning Systems*, 31(8):2734–2749. 1, 2

[Zhu et al., 2024] Zhu, Y., Wang, X., Li, J., Zhang, Y., and Liu, H. (2024). A novel fraud detection method based on machine learning. *Journal of Financial Crime*, 31(1):123–135. 1, 2

Apêndice A

Artigos Publicados

Neste apêndice, o artigo desenvolvido nesta dissertação é apresentado.