MANUTENÇÃO PREDITIVA INTELIGENTE EM UM AMBIENTE DE PRODUÇÃO OPERACIONAL UTILIZANDO DEEP LEARNING E INTERNET DAS COISAS PARA INFERÊNCIA NA PONTA

Matheus Calheira Guimarães de Oliveira[1]; Helaine Pereira Neves[2]; Arthur Gabriel Lima Paim[2]; Robson Carlos Souza Rosario Junior[2]; Bruno Santos Junqueira[2]; Erick Giovani Sperandio Nascimento[2,3]
[1] Bolsista; Iniciação científica - CNPq; matheus.guimaraes@fbter.org.br
[2] Centro Universitário SENAI CIMATEC; Salvador - BA; erick.sperandio@fieb.org.br
[3] Universidade de Surrey; Guildford – Surrey – Inglaterra; erick.sperandio@fieb.org.br

RESUMO

Este Projeto visa aprimorar o estado da arte de modelos de Aprendizagem de Máquina Profunda na área de Manutenção Preditiva Inteligente, explorando novas técnicas e abordagens para classificar falhas em equipamentos rotativos, bem como testar diferentes tipos de microprocessadores embarcados na ponta para inferência que necessitam apenas de computação básica e modelos otimizados. Para tal propósito foram feitos testes de inferências com um modelo de rede neural convolucional em diferentes dispositivos, com o objetivo de adquirir um tempo médio de inferência em cada um deles, de modo com que eles possam eles ser comparados quantitativamente.

PALAVRAS-CHAVE: Microprocessador; Diagnóstico; Prognóstico.

1. INTRODUÇÃO

As técnicas de aprendizagem de máquina profunda (Deep learning - DL) têm proporcionado avanços significativos nas mais diversas áreas do conhecimento, inclusive no diagnóstico e prognóstico de falhas em ativos industriais. É altamente desejável pela indústria poder implementar sistemas inteligentes que possam detectar automaticamente falhas operacionais em estágio inicial e recomendar a intervenção para evitar condições inseguras no processo e no ambiente. No entanto, e apesar dos grandes avanços no desenvolvimento de modelos de DL para a manutenção preditiva inteligente (MPI) de máquinas industriais, disponibilizá-los em ambiente operacional é um desafio que usualmente envolve restrições de processamento, conexões de rede velozes, tráfego de dados, segurança e requisitos de processamento em tempo real, onde geralmente são implementados com dispositivos de Internet das Coisas (IoT) que dependem de onerosos hardwares e com conexões confiáveis sujeitas a restrições da largura da banda, energia e alta latência. Os microprocessadores (MPU), na forma de sistemas ciberfísicos, surgem como solução que preserva a privacidade da informação (já que nenhum dado crítico deixa o dispositivo) e que proporciona a inferência dos modelos na ponta sem latência, já que não necessitam de conexões com a Internet para as decisões em tempo real que afetam seu desempenho. Assim, este projeto tem como desafio aprimorar o estado da arte de modelos de DL na área de MPI, explorando novas técnicas e abordagens para classificar falhas em equipamentos rotativos, bem como testar diferentes tipos de MPUs embarcados na ponta para inferência que necessitam apenas de computação básica e modelos otimizados. Espera-se que esse projeto contribua para o aprimoramento de estudos que visem trazer as pesquisas na área de MPI, DL, e os modelos desenvolvidos, mais próximos da realidade das aplicações industriais e das restrições que usualmente surgem ao se implantar esses tipos de modelos em ambientes reais. O objetivo desses testes foi medir o tempo médio de inferência por amostra de um modelo DL de ponta em diferentes dispositivos de computação para que pudessem ser comparados quantitativamente.

2. METODOLOGIA

Nosso estudo comparou os tempos médios de inferência de uma Rede Neural Convolucional (CNN) para diagnóstico de falhas[1] em várias arquiteturas de hardware, cada uma com uma capacidade de computação diferente. Os testes foram realizados com o Modelo DL PdM-CNN [1], utilizando o Banco de Dados de Falhas em Maquinário (MaFaulDa), desenvolvido pela Universidade Federal do Rio de Janeiro (UFRJ), e o Conjunto de Dados de Rolamentos da Case Western Reserve University (CWRU). Cada um desses bancos de dados consiste principalmente em séries temporais de vibração coletadas de máquinas rotativas sujeitas a diferentes tipos de anomalias em altas taxas de amostragem (variando de 12 kHz a 50 kHz). O objetivo desses testes foi medir o tempo médio de inferência por amostra de um modelo DL de ponta em diferentes dispositivos de computação para que pudessem ser comparados quantitativamente. Os dispositivos de inferência usados

neste estudo foram os seguintes: um laptop com uma GPU NVIDIA GeForce GTX 1650 e uma CPU Intel(R) Core(TM) i5-11300H; um desktop com uma GPU NVIDIA RTX A2000 e uma CPU Intel(R) Core(TM) i5-10400F; e três nós de cluster de computação de alto desempenho (HPC) instalados no Centro de Supercomputação para Inovação Industrial CIMATEC (CS2i), cada um composto por uma GPU NVIDIA A100, uma CPU Intel(R) Xeon(R) Gold 6148 com 40 núcleos e uma GPU NVIDIA Tesla P100-SXM2, respectivamente. Nenhuma sintonia especial de hardware foi realizada para nenhuma dessas arquiteturas de dispositivo. O PdM-CNN foi implantado e testado em sua arquitetura original (ou seja, não foi realizada poda, quantização ou técnica semelhante).

3. RESULTADOS E DISCUSSÃO

Ao analisar os valores de inferência, podemos perceber que o tempo de inferência medido no laptop e no desktop se comporta como se esperaria, levando em consideração sua respectiva potência computacional. Por outro lado, um fenômeno interessante acontece quando analisamos o tempo médio de inferência para os nós de cluster HPC. Ambas as GPUs levam mais tempo para realizar a inferência em cada amostra do que a CPU, mesmo que esta última tenha menos potência computacional. A partir dessa evidência empírica, podemos observar que uma relação direta entre potência computacional e menor tempo de inferência pode não ser garantida o tempo todo. Com essa nova compreensão e realizando tais testes sistemáticos, podemos argumentar que é possível implantar modelos DL em arquiteturas de hardware menos poderosas sem aumentar os tempos de inferência. Além disso, a capacidade de usar dispositivos de computação menos exigentes possibilita o desenvolvimento de hardware menos consumidor de energia e mais econômico para esse tipo de aplicação.

4. CONSIDERAÇÕES FINAIS

A redução do consumo de energia para a realização de inferências de modelos DL é benéfica para o meio ambiente, pois reduz a pegada de carbono. Além disso, dispositivos de aprendizado profundo menos caros são vantajosos porque podem tornar esse tipo de tecnologia mais acessível a um maior número de pessoas em locais onde a infraestrutura cara frequentemente usada para implantar modelos DL não está disponível.

5. REFERÊNCIAS

[1] SOUZA, R. M. et al. Deep learning for diagnosis and classification of faults in industrial rotating machinery. Computers & Industrial Engineering, v. 153, p. 107060, mar. 2021.