Comparando o Desempenho da NPU Intel AI Boost com uma CPU de Propósito Geral

Kenichi Brumati[1], **Igor Freitas**[2], **Arthur F. Lorenzon**[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul – RS – Brasil

[2]Intel, São Paulo – SP – Brasil

{kbrumati,aflorenzon}@inf.ufrgs.br, igor.freitas@intel.com

Resumo. *A crescente demanda por aplicações de Inteligência Artificial tem impulsionado o desenvolvimento de hardware especializado, como as Unidades de Processamento Neural, projetadas para otimizar tarefas de treinamento e inferência. Com sua crescente adoção em processadores modernos e dispositivos embarcados, torna-se essencial avaliar seu impacto no desempenho das aplicações. Assim, neste trabalho, analisamos o desempenho da NPU Intel AI Boost em comparação a um processador de propósito geral , considerando métricas como latência, tempo de inferência e throughput.*

1. Introdução

A crescente demanda por aplicações de Inteligência Artificial (IA) tem impulsionado a busca por *hardware* especializado capaz de lidar com as exigências computacionais de treinamento e inferência dos modelos. Tarefas como reconhecimento de imagens, processamento de linguagem natural e sistemas autônomos exigem alto desempenho aliado à eficiência energética, tornando uma tarefa desafiadora para para processadores de propósito geral tradicionais. Deste modo, para atender tais demandas, novas arquiteturas especializadas têm sido desenvolvidas nos últimos anos, oferecendo otimizações específicas para operações de IA.

Entre essas arquiteturas, as Unidades de Processamento Neural (NPUs – *Neural Processing Units*) vêm ganhando destaque como uma alternativa eficiente para a exeucção de modelos de IA, sendo incorporadas em processadores modernos (e.g., *laptops*) e dispositivos embarcados. Deste modo, com a crescente popularização dessas unidades, torna-se essencial avaliar seu real impacto no desempenho das aplicações, comparando-as com outras arquiteturas computacionais, para determinar quando e como seu uso pode ser vantajoso em relação a outras soluções, como CPUs e GPUs.

Considerando o exposto acima, neste trabalho, investigamos o desempenho da NPU AI Boost da Intel em comparação a um processador de propósito geral (Intel Ultra i7 165U). Resultados experimentais mostram que a NPU foi aproximadamente 10x mais rápida que a CPU na inferência. Essa análise representa um primeiro passo dentro de um estudo mais amplo, que futuramente incluirá a comparação com GPUs e outras arquiteturas especializadas, permitindo um panorama mais abrangente sobre as melhores opções para executar cargas de trabalho de IA.

2. Unidades de Processamento Neural

As NPUs são aceleradores projetados para cargas de trabalho de IA. Diferentemente de CPUs, que possuem arquiteturas de propósito geral para atender a demanda de diversas

aplicações, e de GPUs, que otimizam paralelismo massivo para processamento gráfico e computacional, as NPUs são especializadas em operações matriciais e tensoriais, fundamentais para redes neurais profundas. Assim, a principal característica das NPUs é a presença de unidades vetoriais e *tensor cores*, que permitem processar operações como multiplicação de matrizes e convoluções e forma altamente eficiente. Além disso, as NPUs incorporam técnicas como quantização, que reduz a precisão dos cálculos (e.g., de FP32 para INT8) sem comprometer significativamente a precisão dos modelos, melhorando o desempenho e reduzindo o consumo de energia.

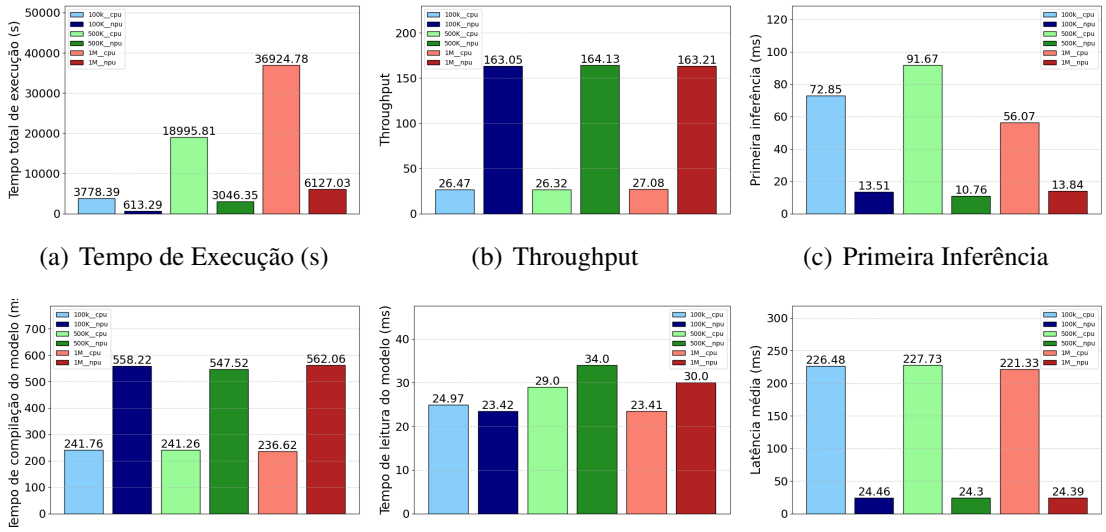
Neste sentido, diversos estudos analisam o desempenho e a arquitetura das NPUs em comparação com outras arquiteturas. O SuperNPU [Ishida et al. 2020] é um simulador para medir desempenho e consumo de energia, enquanto o COOL-NPU combina núcleos CNN-SNN [Kim et al. 2023] e retropropagação orientada por eventos para maior eficiência em aprendizado online. Outras pesquisas focam na virtualização e no acesso à memória. Em [Xue et al. 2023], são exploradas técnicas para otimizar o uso de NPUs em ambientes de nuvem, enquanto [Zhang et al. 2019] analisa o impacto das memórias resistivas no processamento neural. Comparando NPUs com outros processadores, [Wu et al. 2022] discute aceleradores especializados para inferência e os desafios na construção de hardware eficiente para IA. Assim, este trabalho complementa essas análises ao avaliar o desempenho da NPU Intel em relação a um processador de propósito geral, estabelecendo um primeiro passo para futuras comparações com GPUs e outras arquiteturas especializadas.

3. Metodologia

Os experimentos foram realizados em um laptop Lenovo Mobile com processador Intel Ultra i7 165U, 16GB DDR5 RAM e arquitetura Meteor Lake, com Windows 11. O equipamento conta com a NPU integrada Intel AI Boost, compatível com os *frameworks* OpenVINO, WindowsML, DirectML e ONNX RT. Para avaliar o desempenho da NPU em relação à CPU, utilizou-se OpenVINO 2024.6.0 e Python 3.11. A ferramenta permite a execução de testes selecionando diferentes modelos de IA, sendo escolhidos *Resnet-50-tf* e *Mobilenet* com integração YOLO v4. Os testes foram realizados com 100k, 500k e 1M inferências, coletando métricas como tempo de leitura e compilação do modelo, tempo da primeira inferência, tempo total de execução, latências (média, máxima e mínima) e *throughput*. A análise focou nos impactos no tempo total de execução, *throughput* e tempo de compilação.

4. Resultados Experimentais

Nesta seção são discutidos os resultados de desempenho da execução dos dois modelos descritos na Metodologia nas arquiteturas alvo. Para isso, a Figura 1 apresenta os resultados obtidos na execução do modelo *Mobilenet*, enquanto que a Figura 2 mostra os resultados do modelo *Resnet*. Dentre os resultados coletados, observa-se a clara superioridade da NPU no processamento de cálculos. Nos testes realizados, a execução das operações apresentou uma certa linearidade. Ao comparar ambos os modelos (*Mobilenet* e *Resnet*), nota-se uma diferença evidente no tempo de execução, que se torna mais acentuada à medida que o número de operações aumenta. No caso do modelo *Mobilenet* (Figura 1a), ao executar um milhão de iterações, a CPU levou 36.924,78 segundos, enquanto a NPU completou a tarefa em 6.123,03 segundos, resultando em um ganho de 6,03 vezes. Já para o modelo *Resnet* (Figura 2a), a NPU obteve um ganho ainda maior, sendo 10,05 vezes mais rápida que a CPU. Esses resultados sugerem que, dependendo da



(a) Tempo de Execução (s) (b) Throughput (c) Primeira Inferência

(d) Tempo de Compilação (e) Tempo de Leitura do Modelo (f) Latência Média

Figura 1. Resultados de desempenho para execução do modelo *mobilenet*.

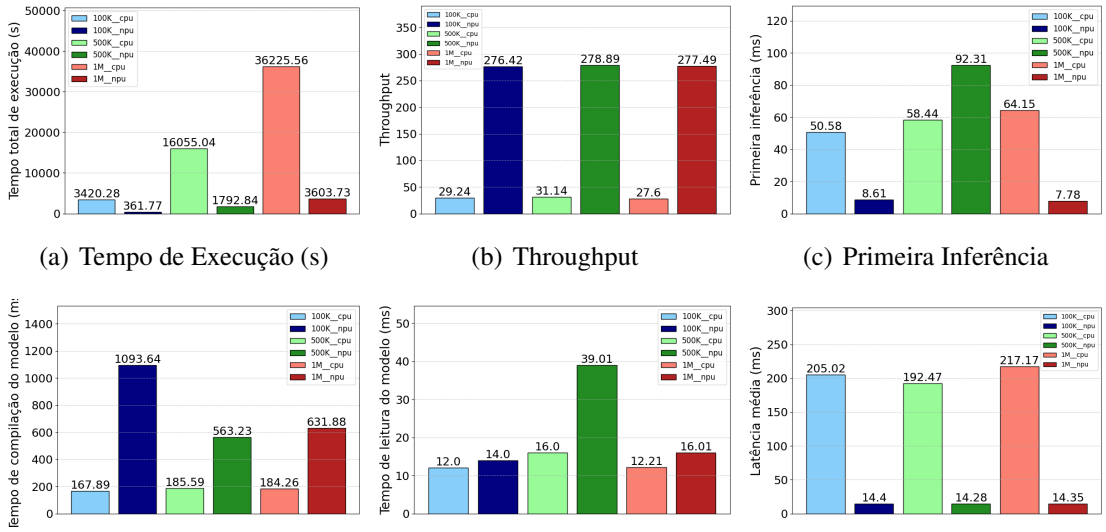
complexidade dos cálculos, o uso de hardware especializado pode ser significativamente mais eficiente, especialmente para operações mais simples, onde a precisão não é crítica.

Outro ponto importante é o *throughput* do modelo, em que ambos os modelos, a NPU atingiu o poder de processamento maior, se mantendo praticamente o mesmo ao variar o modelo e número de inferências, já que ele é calculado pelo número de operações durante um determinado tempo, podendo haver pequenas variações de tempo entre as execuções de cálculos resultando nessa diferença. Em ambos os modelos de IA testados a NPU teve um ganho na capacidade de processamento, que varia de 6,15 à 9,46 vezes comparado a CPU. Referente a velocidade em que foram realizadas cada inferência, destaca-se, a NPU processando muito mais rápido que a CPU, o que ajuda a confirmar os dados expostos anteriormente quando a diferença no tempo de execução.

Quando analisamos o tempo de leitura do modelo, ambos os processadores apresentaram valores semelhantes na maioria dos casos. No entanto, no processo de compilação, a CPU demonstrou maior velocidade. Isso sugere que, para modelos de IA mais complexos, a compilação diretamente na NPU pode apresentar dificuldades. Assim, uma abordagem mais eficiente seria realizar a compilação externamente e transferir apenas as informações necessárias para o cálculo. Em relação à latência média, a NPU apresenta valores mais baixos, pois é ativada apenas quando solicitada para processamento, garantindo maior eficiência nesse aspecto. Já a CPU, por executar diversas outras operações simultaneamente, não consegue dedicar todos os seus recursos exclusivamente ao processamento da IA, o que pode resultar em uma latência superior.

5. Conclusão

Os experimentos realizados demonstraram a superioridade da NPU Intel AI Boost em relação à CPU Intel Ultra i7 165U, especialmente em tarefas de inferência, onde a NPU alcançou até 10 vezes mais desempenho. Além disso, a NPU apresentou menor latência e maior *throughput*, confirmando sua eficiência para cargas de trabalho de IA. No entanto, a CPU demonstrou vantagem no tempo de compilação, sugerindo que modelos mais complexos podem se beneficiar de uma abordagem híbrida, realizando a compilação externamente antes da execução na NPU. Esses resultados representam um primeiro passo



(a) Tempo de Execução (s) (b) Throughput (c) Primeira Inferência

(d) Tempo de Compilação (e) Tempo de Leitura do Modelo (f) Latência Média

Figura 2. Resultados de desempenho para execução do modelo *resnet*.

na avaliação de NPUs, abrindo caminho para futuras comparações com GPUs e outras arquiteturas especializadas.

Agradecimentos

Os autores gostariam de agradecer a Intel Brasil pelo apoio a pesquisa provendo acesso ao hardware necessário para conduzir este trabalho e também a FAPERGS, CAPES e CNPq pelo financiamento a pesquisa.

Referências

Ishida, K., Byun, I., Nagaoka, I., Fukumitsu, K., Tanaka, M., Kawakami, S., Tanimoto, T., Ono, T., Kim, J., and Inoue, K. (2020). Supernpu: An extremely fast neural processing unit using superconducting logic devices. In *Annual IEEE/ACM MICRO*, pages 58–72.

Kim, S., Kim, S., Hong, S., Kim, S., Han, D., Choi, J., and Yoo, H.-J. (2023). Cool-npu: Complementary online learning neural processing unit with cnn-snn heterogeneous core and event-driven backpropagation. In *IEEE COOL CHIPS*, pages 1–3.

Wu, B., Furtner, W., Waschneck, B., and Mayr, C. (2022). Industry-track: Towards agile design of neural processing unit. In *CODES+ISSS*, pages 17–20.

Xue, Y., Liu, Y., and Huang, J. (2023). System virtualization for neural processing units. In *HOTOS*, page 80–86.

Zhang, W., Peng, X., Wu, H., Gao, B., He, H., Zhang, Y., Yu, S., and Qian, H. (2019). Design guidelines of rram based neural-processing-unit: A joint device-circuit-algorithm analysis. In *ACM/IEEE DAC*, pages 1–6.