Avaliação Experimental dos Trade-offs entre Precisão Numérica, Desempenho e Eficiência Energética em Inferência com TensorRT

Murilo Salem[1], **Daniel Pontes**[1], **Luísa Bohm**[1], **Henrique dos Reis**[1], **Gerson Geraldo H. Cavalheiro**[1]

[1]Centro de Desenvolvimento Tecnológico
Universidade Federal de Pelotas
96.010-610 – Pelotas – RS – Brazil

{mcsalem,dhspbarretos,lhbohn,hdreis,gersonc}@inf.ufpel.edu.br

Resumo. *A inferência eficiente de modelos de aprendizado profundo em GPU depende fortemente da representação numérica adotada. Este trabalho avalia os efeitos de FP16, BF16 e INT8 sobre acurácia, latência, throughput, energia por amostra e tamanho do engine, tomando FP32 como baseline, em inferência com ResNet-50, ImageNet-100, TensorRT e* `trtexec` *em uma NVIDIA GeForce RTX 5090. A campanha principal gerou 364 registros brutos, e a análise inferencial considerou 30 repetições por precisão nos lotes 1, 8 e 32. Os resultados mostram que INT8 obteve o maior desempenho bruto e a melhor eficiência energética, mas com perda de 2,92 p.p. em Top-1 em relação ao FP32. Sob o critério de aceitação adotado — speedup* $\geq 1{,}15\times$ *e* Δ*Top-1* $\geq -1{,}0$ *p.p. — FP16 e BF16 foram classificadas como vantajosas em todos os tamanhos de lote. Entre elas, FP16 apresentou o melhor compromisso global, com speedup entre* $4{,}22\times$ *e* $4{,}54\times$ *e redução energética entre 70,5% e 78,4%, sem perda observável de Top-1 nesta campanha. A análise inferencial confirmou diferenças estatisticamente significativas entre as precisões para latência, throughput e energia em todos os lotes avaliados (*$p < 0{,}05$*).*

1. Introdução

A inferência de modelos de aprendizado profundo em GPU tornou-se um componente central em sistemas modernos de visão computacional, processamento de linguagem natural e aplicações industriais. Nesse contexto, reduzir latência, elevar *throughput* e melhorar eficiência energética sem comprometer significativamente a qualidade preditiva é um problema relevante para computação de alto desempenho [Reddi et al. 2020].

Uma estratégia amplamente empregada para esse fim consiste na adoção de representações numéricas mais compactas, como *FP16*, *BF16* e *INT8*. Em princípio, tais formatos permitem reduzir custo computacional e footprint do modelo, mas o efeito real depende do modelo, do *runtime*, do hardware e do procedimento de quantização adotado. Assim, comparações objetivas e reproduzíveis permanecem necessárias [Micikevicius et al. 2018, Jacob et al. 2018, Reddi et al. 2020].

Este trabalho investiga como *FP16*, *BF16* e *INT8* afetam acurácia, latência, *throughput*, memória, energia e tamanho do *engine* na inferência de *ResNet-50* com *TensorRT*, tomando *FP32* como referência. As principais contribuições são: (i) um bench-

mark reproduzível baseado em *ONNX*+*TensorRT*+`trtexec`; (ii) um pipeline automatizado de agregação estatística descritiva e geração de relatórios; e (iii) uma etapa de análise inferencial com testes de normalidade e comparação par-a-par entre precisões.

2. Metodologia

O benchmark foi conduzido em um único nó com uma GPU NVIDIA GeForce RTX 5090, driver 590.48.01, Python 3.12.3, CUDA 12.8, PyTorch 2.10.0+cu128, torchvision 0.25.0+cu128 e TensorRT 10.15.1.29. O modelo avaliado foi *ResNet-50* com pesos `DEFAULT` do torchvision [He et al. 2016]. O conjunto de dados utilizado foi o *ImageNet-100*, com 5000 amostras de validação para a fase de acurácia e 500 amostras de treino representativas para a calibração *INT8*.

As configurações comparadas foram *FP32* (baseline), *FP16*, *BF16* e *INT8*. O fluxo experimental adotou exportação para ONNX, construção de *engines* com *TensorRT* e medição de desempenho com `trtexec`. No caso de *INT8*, foi utilizada quantização pós-treinamento estática com calibração baseada em dados representativos de treino [NVIDIA 2026].

A campanha principal considerou os tamanhos de lote 1, 8 e 32, com 30 repetições por precisão para a análise inferencial. Cada repetição executou 100 iterações de *warm-up* e 1000 iterações de medição via `trtexec`. A fase de acurácia foi realizada separadamente da fase de desempenho.

As métricas coletadas foram: acurácia Top-1 e Top-5, latência média, latência p50, latência p95, *throughput* em amostras por segundo, pico de memória de GPU, energia por amostra e tamanho do *engine*. Foram computadas estatísticas descritivas e intervalos de confiança de 95%. Uma configuração foi considerada *vantajosa* quando apresentou *speedup* $\geq 1{,}15$ em relação ao *FP32* e ΔTop-1 $\geq -1{,}0$ p.p. em relação ao baseline.

Adicionalmente, foi conduzida análise estatística inferencial sobre latência média, *throughput* e energia por amostra. Para cada grupo, a normalidade foi verificada com Shapiro-Wilk. As seis comparações par-a-par entre precisões foram então realizadas com seleção automática entre *Welch t-test* e *Mann-Whitney U*, conforme as premissas observadas. ANOVA *one-way* com pós-teste de Tukey HSD foi configurada como análise opcional, sendo executada apenas quando as premissas de normalidade e homogeneidade de variâncias eram satisfeitas.

Os dados e scripts do benchmark serão disponibilizados em repositório GitHub público em caso de aprovação do artigo.

3. Resultados

A Tabela 1 resume os principais resultados obtidos na campanha principal. O *FP32* foi usado como referência estrita. Entre as configurações avaliadas, *INT8* apresentou o maior desempenho bruto, a menor latência média, a melhor eficiência energética e o menor tamanho de *engine*, porém com perda de 2,92 p.p. em Top-1 em relação ao baseline, ultrapassando o limiar de fidelidade adotado no protocolo. Já *FP16* e *BF16* satisfizeram simultaneamente os critérios de *speedup* e acurácia, sendo ambas classificadas como vantajosas.

Sob a ótica do compromisso prático entre fidelidade e eficiência, *FP16* apresentou o melhor resultado global nesta campanha, com *speedup* entre $4{,}22\times$ e $4{,}54\times$ e redução energética entre 70,5% e 78,4%. Embora *BF16* também tenha sido vantajosa em todos os tamanhos de lote, permaneceu atrás de *FP16* em *throughput* e energia.

Em acurácia Top-1, *FP16* e *BF16* apresentaram valores superiores ao *FP32* nesta campanha, ao passo que *INT8* registrou queda de 2,92 p.p. Esse resultado deve ser interpretado como ausência de degradação observável para *FP16* e *BF16* no protocolo adotado, e não como evidência de que formatos de menor precisão aumentem inerentemente a capacidade preditiva do modelo.

Table 1. Resumo consolidado dos resultados da campanha principal. Valores positivos de ΔTop-1 indicam ganho observado em relação ao FP32; valores negativos indicam perda observada.

Precisão	Top-1 (%)	Speedup (1/8/32)	Redução de energia (%)	*Engine* (MB)	ΔTop-1 vs FP32 (p.p.)	Status
FP32	46.44	1.00 / 1.00 / 1.00	0.0	98.19	0.00	baseline
FP16	47.24	4.22 / 4.54 / 4.51	70.5–78.4	49.27	+0.80	vantajosa
BF16	48.70	3.45 / 3.01 / 2.92	66.6–70.3	49.30	+2.26	vantajosa
INT8	43.52	5.22 / 5.27 / 5.26	77.9–83.2	25.65	-2.92	não vantajosa

O padrão permaneceu estável nos três tamanhos de lote: *INT8* dominou desempenho bruto, enquanto *FP16* ofereceu a melhor combinação entre desempenho e fidelidade dentro do critério adotado.

A análise inferencial reforçou que as diferenças observadas entre precisões não se explicam por ruído amostral. Para latência média, *throughput* e energia por amostra, todas as seis comparações par-a-par entre *FP32*, *FP16*, *BF16* e *INT8* foram estatisticamente significativas em todos os tamanhos de lote avaliados ($p < 0{,}05$). A seleção entre *Welch t-test* e *Mann-Whitney U* foi realizada automaticamente com base nas premissas observadas em cada caso. Embora ANOVA *one-way* com Tukey HSD tenha sido prevista no protocolo, ela foi omitida nos nove cenários analisados devido à recorrente violação das premissas de normalidade residual e, principalmente, de homogeneidade de variâncias.

Outro resultado relevante foi a redução do tamanho do *engine*. Em comparação ao *FP32*, os *engines FP16* e *BF16* ocuparam aproximadamente metade do espaço, enquanto o *INT8* reduziu esse tamanho para cerca de um quarto. Em contraste, a telemetria de pico de memória de GPU mostrou diferenças relativamente pequenas entre as configurações, sugerindo que essa métrica capturou majoritariamente o overhead global do *runtime*, e não apenas o *footprint* efetivo do modelo.

4. Discussão e ameaças à validade

Os resultados mostram que a escolha da representação numérica altera substancialmente o compromisso entre fidelidade e eficiência. No cenário investigado, *FP16* apresentou o melhor compromisso global sob a restrição de acurácia adotada, enquanto *BF16* permaneceu vantajosa e *INT8* dominou desempenho bruto ao custo de degradação preditiva acima do limiar aceito.

A análise inferencial reforça essa leitura ao mostrar que as diferenças de latência, *throughput* e energia entre precisões foram estatisticamente significativas em todos os

tamanhos de lote analisados. Ao mesmo tempo, a inviabilidade recorrente da ANOVA devido à heterogeneidade de variâncias e à não normalidade residual reforça a adequação de um protocolo que combine testes paramétricos e não paramétricos de forma adaptativa.

Este estudo apresenta, contudo, algumas ameaças à validade. Primeiro, o benchmark principal foi executado apenas com *ResNet-50*, o que limita a generalização para outras arquiteturas. Segundo, a avaliação foi conduzida sobre *ImageNet-100* com mapeamento para o espaço de classes do *ImageNet-1K*; assim, os valores absolutos de acurácia devem ser interpretados com cautela. Terceiro, as métricas de energia e memória foram obtidas por *polling* com `nvidia-smi`, e não por instrumentação de alta frequência.

5. Conclusão

Este trabalho apresentou um benchmark reproduzível para avaliar o impacto de *FP16*, *BF16* e *INT8* na inferência de *ResNet-50* com *TensorRT*, incorporando tanto sumarização estatística descritiva quanto análise inferencial. No cenário avaliado, *INT8* apresentou o melhor desempenho bruto e a maior eficiência energética, mas com perda de acurácia acima do limiar definido no protocolo. Entre as configurações que satisfizeram simultaneamente os critérios de desempenho e fidelidade, *FP16* foi a que apresentou o melhor compromisso global, com *speedup* entre $4{,}22\times$ e $4{,}54\times$ e redução energética entre 70,5% e 78,4%, sem perda de Top-1 nesta campanha. *BF16* também permaneceu vantajosa em todos os lotes testados, embora com desempenho inferior ao *FP16*. Além disso, a análise inferencial mostrou que as diferenças entre precisões para latência, *throughput* e energia foram estatisticamente significativas em todos os tamanhos de lote considerados. Como trabalho futuro, pretende-se ampliar o estudo para *DistilBERT*, investigar comparações entre *post-training quantization* e *quantization-aware training*, e estender a avaliação estatística para múltiplas arquiteturas e cargas de trabalho.

References

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., and Kalenichenko, D. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2704–2713. IEEE.

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., and Wu, H. (2018). Mixed precision training. In *International Conference on Learning Representations (ICLR)*.

NVIDIA (2026). Command-line programs — nvidia tensorrt documentation. `https://docs.nvidia.com/deeplearning/tensorrt/latest/reference/command-line-programs.html`. Acesso em: 6 mar. 2026.

Reddi, V. J., Cheng, C., Kanter, D., Mattson, P., Schmuelling, G., Wu, C.-J., Anderson, B., Khailo, M., Jan, J.-W., Esmaeilzadeh, H., et al. (2020). Mlperf inference benchmark. In *Proceedings of the 2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*, pages 446–459. IEEE.