Relatório de Benchmark — Inferência com NVIDIA Triton para Detecção de Armas

Este relatório apresenta os resultados de um benchmark local utilizando o **NVIDIA Triton Inference Server** para o modelo de **detecção de armas** baseado em YOLO.

Descrição do Modelo

- **Arquitetura**: YOLOv11 Large
- **Conversão**: Modelo convertido para **TensorRT** para acelerar inferência
- **Otimização**:
  - Suporte a **batch size dinâmico**
  - Precisão e otimizações específicas para a **GPU NVIDIA**
- **Gerenciamento de Concorrência**: Delegado ao **NVIDIA Triton**, que distribui múltiplas requisições simultâneas de inferência.

Ambiente de Execução Local

- **Máquina**: Notebook Dell G15
- **GPU**: NVIDIA RTX 3050 Laptop GPU (6 GB VRAM)
- **Sistema**: Ambiente local com servidor Triton configurado para aceitar diferentes níveis de concorrência e batch size.

1. Throughput vs Concorrência

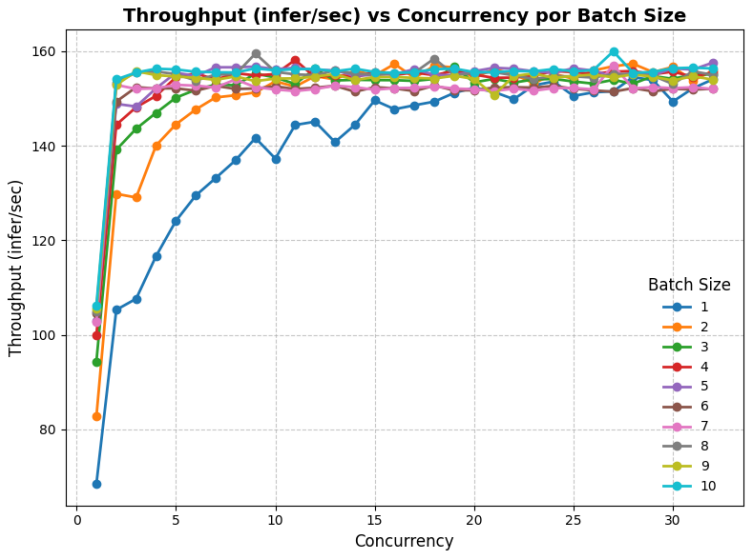

Figure 1: Throughput vs Concurrency

- O throughput atinge o máximo com batch sizes ≥ 4 e concorrência ≥ 10.
- Após esse ponto, o ganho é marginal.
- Excelente escalabilidade com aumento inicial de concorrência.

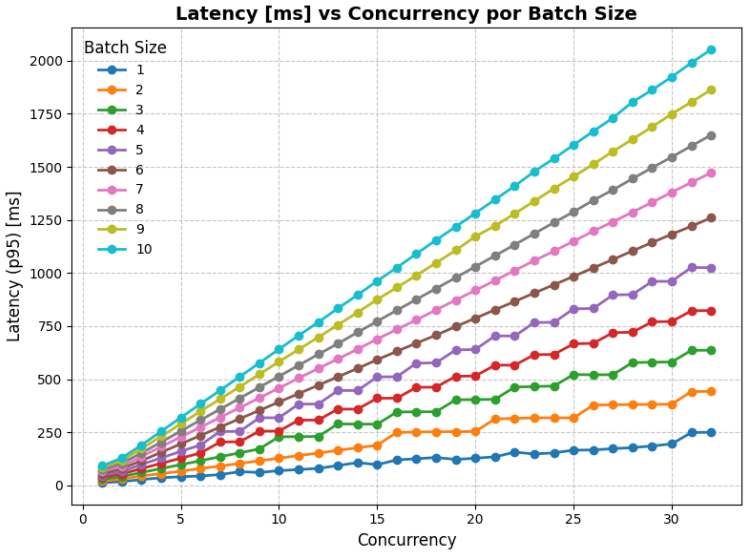

Figure 2: Latency vs Concurrency

2. Latência (p95) vs Concorrência

- A latência cresce de forma aproximadamente linear.
- Batch sizes maiores aumentam a latência significativamente.
- Para aplicações em tempo real, batch size = 1 ou 2 é recomendado.

3. Utilização Média da GPU

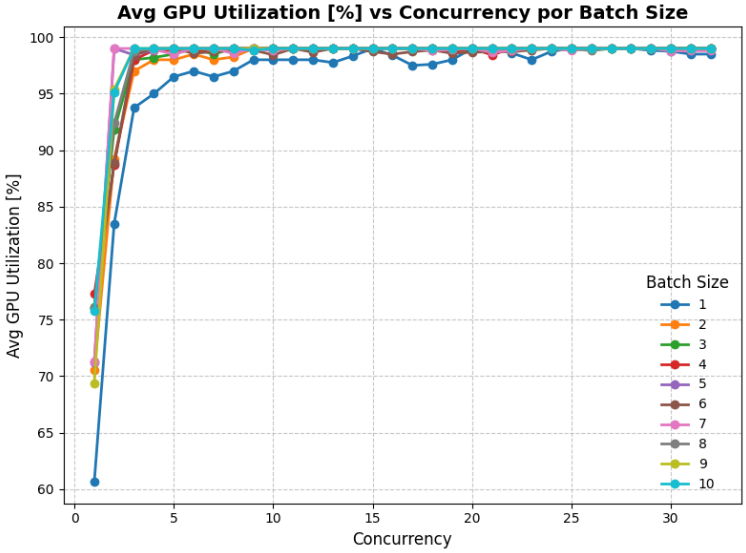

Figure 3: GPU Utilization vs Concurrency

- A GPU rapidamente atinge utilização próxima de 100% com poucos clientes simultâneos.
- O modelo explora bem os recursos da RTX 3050 mesmo em ambiente local.

4. Uso Máximo da Memória da GPU

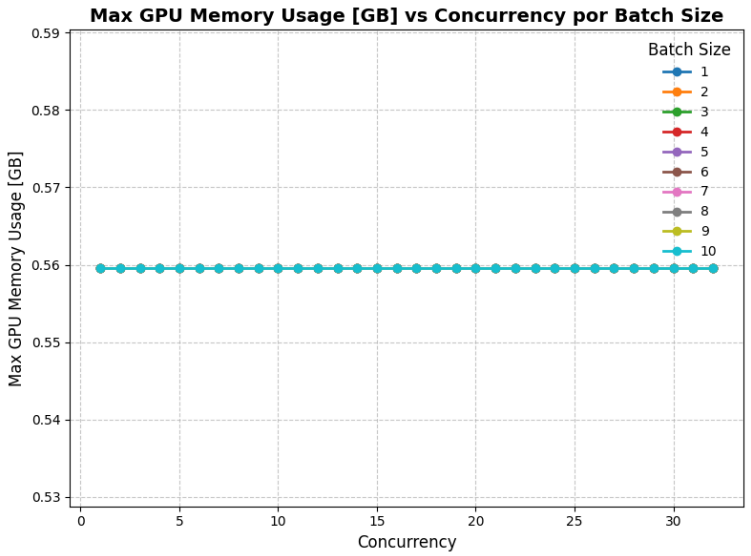

Figure 4: Max GPU Memory Usage

- O uso de memória é extremamente eficiente e estável (~0.56 GB).
- Indica ótima viabilidade para implantação em ambientes com recursos limitados.

Conclusões (Ambiente Local)

- **Melhor throughput**: `batch size $\geq$ 4` com `concorrência $\geq$ 10`.
- **Menor latência**: `batch size = 1 ou 2` com `concorrência  8`.
- **Alta eficiência da GPU**: Excelente aproveitamento da RTX 3050.
- **Memória estável**: Sem variação com o batch size ou concorrência.

[A SER COMPLETADO] Benchmark no Servidor da T4S

Este espaço será preenchido com os resultados de benchmark no ambiente de servidor dedicado com GPU(s) de alto desempenho (L40), visando validar a escalabilidade do modelo em infraestrutura de produção.

Recomendação

Para o modelo de detecção de armas via YOLOv11-Large com Triton:

Objetivo	Batch Size	Concorrência	Comentário
Baixa Latência	1–2	1–8	Ideal para detecções em hardware embarcado
Alto Throughput	4–6	≥ 10	Ideal para processamento em lote ou remoto
Uso eficiente de recursos	≥ 2	≥ 5	Boa utilização da GPU e memória

Objetivo	Batch Size	Concorrência	Comentário
Escalabilidade	A ser testado no servidor dedicado		

Relatório gerado com base em testes locais utilizando o NVIDIA Triton Inference Server em ambiente controlado.