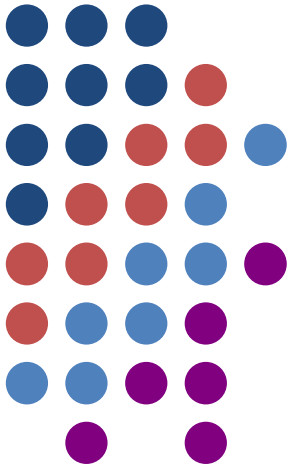
中国科学院高能物理研究所
Institute of High Energy Physics
Chinese Academy of Sciences

八月季度考核报告

学生：肖鹏飞
导师：季筱璐
2024年8月30日



主要内容

- TNS论文撰写投稿
- HEPS-BPIX4 DAQ基于机器学习目标检测的在线图像数据压缩研究
  - HEPS-BPIX4 DAQ在线数据压缩
  - YOLOv10实时端到端的目标检测模型
  - 数据集准备与模型训练
  - 训练模型的推理与图像压缩测试
  - HEPS-BPIX4 DAQ框架部署测试
- 当前计划

TNS论文撰写投稿

- P. Xiao, X. Ji and P. Cao “Research for Application of SIMD/SIMT-Based Online Processing Acceleration Technology.”

Research for Application of SIMD/SIMT-Based Online Processing Acceleration Technology

Pengfei Xiao, Xiaolu Ji, and Ping Cao

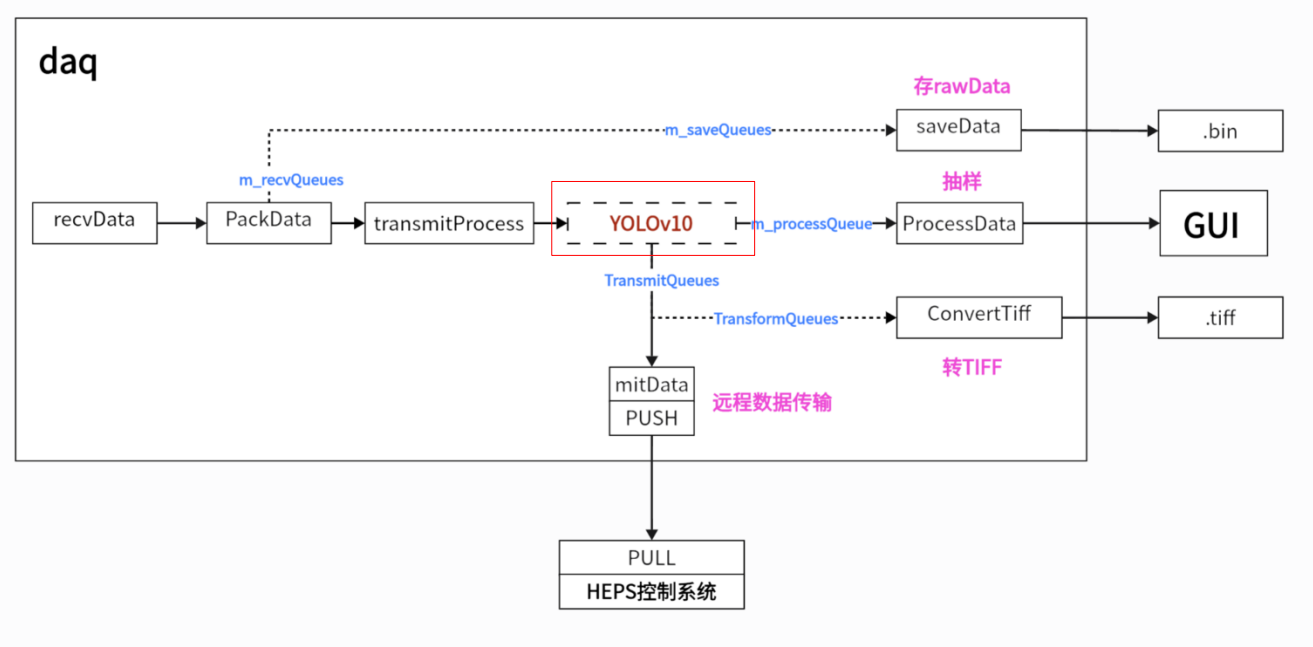
***Abstract*—Based on the challenge of the increasing amount of data that need to be processed online in the current high-energy physics experiments, a high-performance online data processing method is urgently needed. It is a feasible scheme to accelerate and optimize the current online data processing to meet the performance requirements of the online processing process under large amount of data. SIMD (Single Instruction Multiple Data) and SIMT (Single Instruction Multiple Data) are parallel acceleration technologies based on the characteristics of hardware architecture. They can realize the parallel computing processing of online data based on CPU and GPU, and have great application potential for performance improvement. In this paper, according to some data processing requirements of the current actual physical experiment, the performance optimization of the data processing process is implemented based on CPU and GPU, and the related acceleration methods are studied and tested. By exploring the application potential of SIMD and SIMT technologies in online data processing of physical experiments, this paper aims to provide effective scheme reference and solution ideas for online data accelerated processing of future experiments.**

multiplication on a group of data in a single operation. Common SIMD instruction sets include Intel's SSE (Streaming SIMD Extensions) and AVX (Advanced Vector Extensions), and AMD's SSE and AVX. The SIMT (Single Instruction Multiple Thread) technology on GPU is actually a superset of SIMD [4]. The batch processing of data can also be realized by enabling multithreading on GPU. Because there are more computing units in GPU, the parallel computing ability is stronger, and it can be applied to more computing scenarios. Based on the above characteristics, the implementation of SIMD and SIMT on CPU and GPU can effectively accelerate the intensive computing process in experiments.

The research results can provide references for the general data processing methods based on heterogeneous computing in physical experiments, and it is expected that on the basis of this research, various computing platforms could be integrated to build a general high-performance heterogeneous computing

HEPS-BPIX4 DAQ基于机器学习目标检测的在线图像数据压缩研究

- **HEPS-BPIX4 DAQ在线数据压缩**

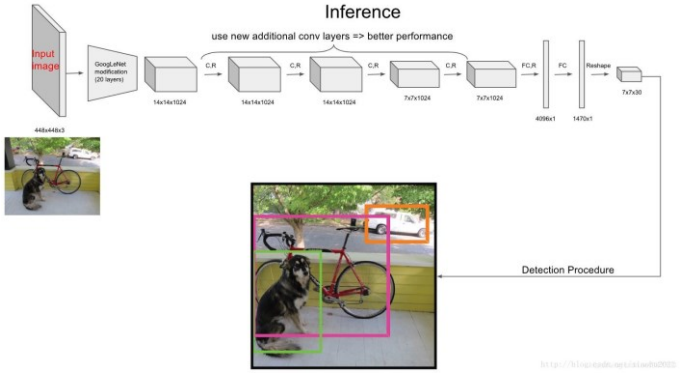


- 在原本的在线数据流框架中对图像数据进行目标区域的检测选取，以减小图像数据的大小实现图像数据的压缩，减轻后续数据处理和传输的压力
- 选用YOLOv10目标检测模型进行训练并于框架中进行部署和集成

HEPS-BPIX4 DAQ基于机器学习目标检测的在线图像数据压缩研究

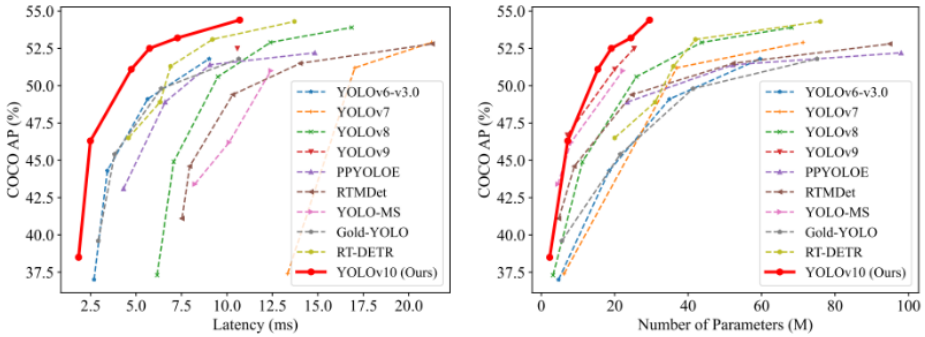
- **YOLOv10实时端到端的目标检测模型**

YOLO 推理结构示意图



端到端的目标检测通过一次前向传播就可以完成检测图形目标的位置和类别

YOLOv10 性能对比



与YOLOv9相比，在相同性能下，YOLOv10的延迟减少了46%，参数减少了25%

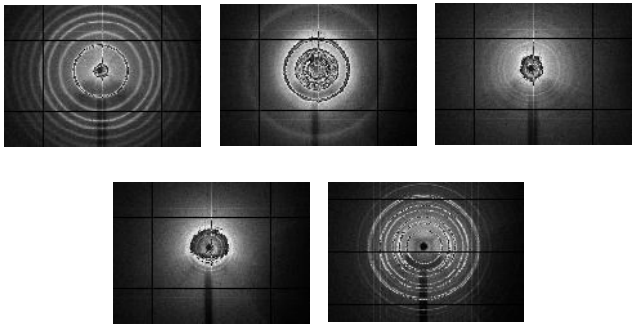
- YOLO实时目标检测算法的模型是一个轻量级的卷积神经网络，其目的是为了解决传统目标检测方法速度慢、效率低的问题
- YOLOv10模型于2024年5月23日发布，致力于打造实时端到端的高性能目标检测器，推理实时性相比于前代模型有较大的提升，贴合在线处理中对实时性的要求

HEPS-BPIX4 DAQ基于机器学习目标检测的在线图像数据压缩研究

- 数据集准备与模型训练

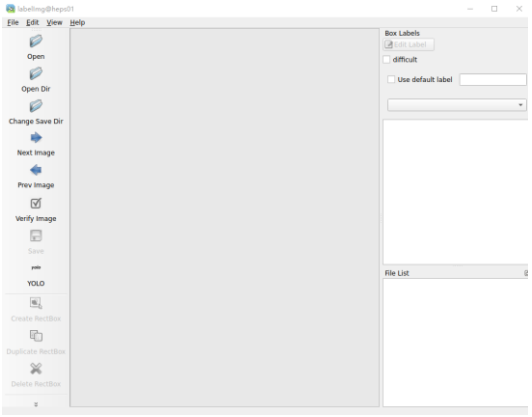
YOLO数据集标注示意图

待标注数据集



导入

LabelImg工具数据标注



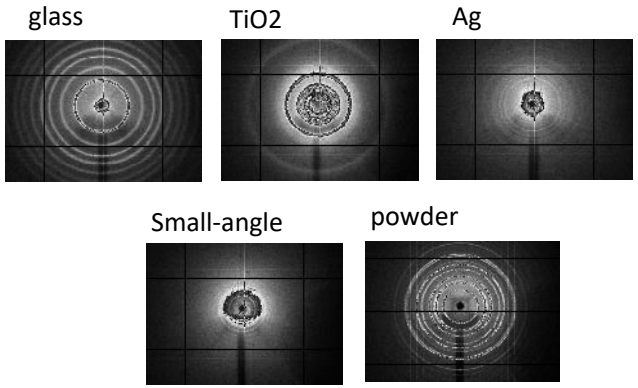
数据集标注与分类（将不同类别的数据混合制作数据集，标注目标区域位置）

- 数据集标注的质量影响后续模型训练的效率和推理的准确性
- 标注过程主要为手工标注，自动标注工具需要提前导入已训练的模型（本研究未采用）

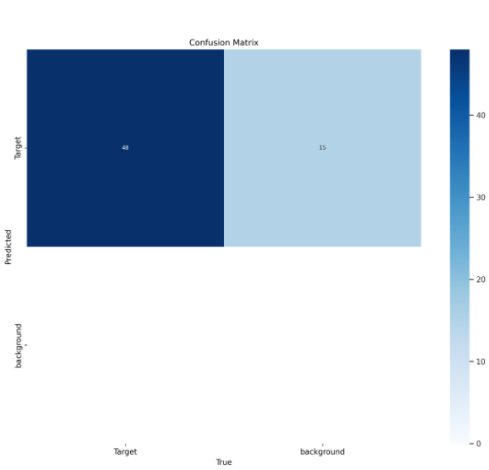
HEPS-BPIX4 DAQ基于机器学习目标检测的在线图像数据压缩研究

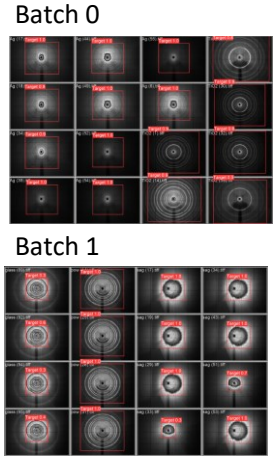
- 数据集准备与模型训练

数据集分类与标注



YOLO模型训练结果示意图





标注324张.tiff图片(TiO2/glass/Ag/small-angle/powder)生成324个标签数据集按Yolo格式进行划分(训练集：226；验证集：48；测试集：50)

混淆矩阵

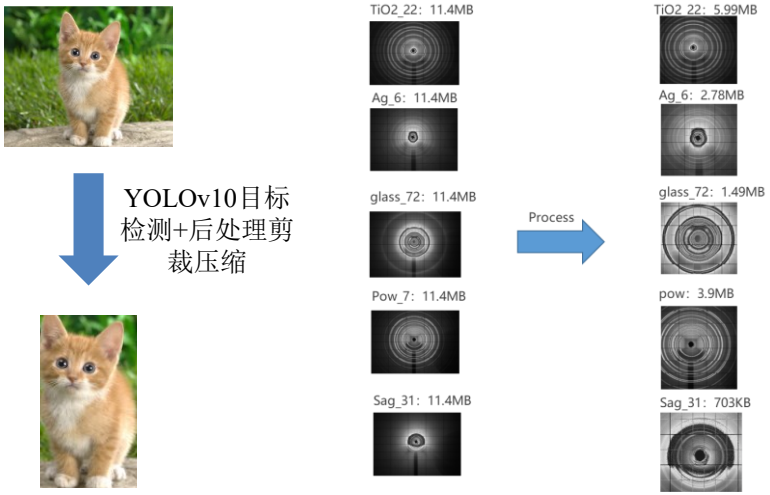
验证集测试

- 训练 A40单卡进行训练（轮次：100；单批训练集大小：16；平均利用率：83%）
- 验证集48个数据的验证结果（识别48，未识别0），识别率100%

HEPS-BPIX4 DAQ基于机器学习目标检测的在线图像数据压缩研究

- 训练模型的推理与图像压缩测试

图像目标检测与压缩



推理性能测试

CPU部署测试

Latency/ms	Ag_6	Glass_72	Pow_7	Sag_31	TiO2_22
Preprocess	6.7	17.4	17.5	18.5	17.6
Inference	28	31.7	27.1	32.0	26.9
Postprocess	1.3	0.6	0.7	1.1	0.7
Utilization	251.3%	223.4%	241.3%	427.4%	407.9%

GPU部署测试

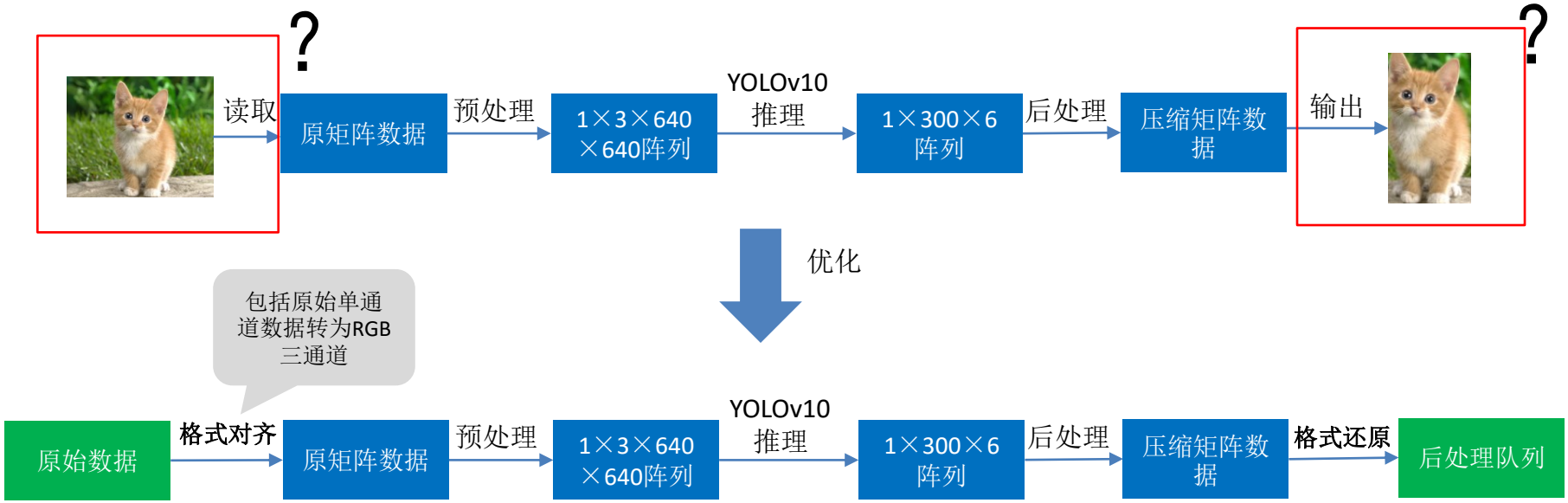
Latency/ms	Ag_6	Glass_72	Pow_7	Sag_31	TiO2_22
Preprocess	3.6	3.8	3.7	3.6	3.6
Inference	3.7	3.7	3.7	3.8	3.6
Postprocess	0.9	0.9	0.9	1.0	0.9
Utilization	79%	73%	100%	60%	65%

- 修改模型的图像后处理逻辑，实现输入图像的剪裁压缩
- Intel Xeon 8462Y+ CPU部署测试的单张图像的处理平均时延为45.56ms，平均利用率为310.26%；Nvidia A40 GPU部署测试的单张图像的处理平均时延为8.28ms，平均利用率为75.4%。平均压缩73.95%

HEPS-BPIX4 DAQ基于机器学习目标检测的在线图像数据压缩研究

- **HEPS-BPIX4 DAQ框架部署测试**

HEPS-BPIX4 DAQ数据流部署设计

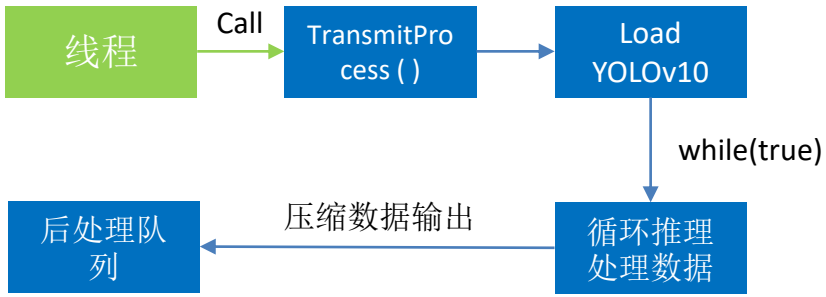


- 实际省略图像生成和读取过程直接将原始数据进行格式对齐输入模型进行推理以提升整体流程性能
- 为了更好地集成如DAQ框架，利用OpenCV库与onnxruntime跨平台引擎用C++重构训练模型的部署

HEPS-BPIX4 DAQ基于机器学习目标检测的在线图像数据压缩研究

- **HEPS-BPIX4 DAQ框架部署测试**

HEPS-BPIX4 DAQ数据流部署示意

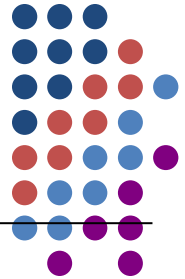


DAQ框架中单线程处理运行

```
initialize ok!!
warning!!! the headFlag is not 0x55aa, headFlag:45ac
duration = 901ms
inference success!!!
warning!!! the headFlag is not 0x55aa, headFlag:45ac
duration = 9ms
inference success!!!
warning!!! the headFlag is not 0x55aa, headFlag:45ac
warning!!! the headFlag is not 0x55aa, headFlag:cdac
duration = 8ms
inference success!!!
warning!!! the headFlag is not 0x55aa, headFlag:45ac
warning!!! the headFlag is not 0x55aa, headFlag:cdac
duration = 8ms
inference success!!!
warning!!! the headFlag is not 0x55aa, headFlag:45ac
duration = 7ms
inference success!!!
warning!!! the headFlag is not 0x55aa, headFlag:45ac
duration = 7ms
inference success!!!
warning!!! the headFlag is not 0x55aa, headFlag:45ac
duration = 7ms
inference success!!!
warning!!! the headFlag is not 0x55aa, headFlag:45ac
duration = 6ms
inference success!!!
warning!!! the headFlag is not 0x55aa, headFlag:45ac
duration = 7ms
inference success!!!
warning!!! the headFlag is not 0x55aa, headFlag:45ac
duration = 7ms
inference success!!!
duration = 7ms
inference success!!!
duration = 6ms
inference success!!!
warning!!! the headFlag is not 0x55aa, headFlag:45ac
warning!!! the headFlag is not 0x55aa, headFlag:45ac
duration = 7ms
inference success!!!
warning!!! the headFlag is not 0x55aa, headFlag:45ac
warning!!! the headFlag is not 0x55aa, headFlag:45ac
duration = 6ms
inference success!!!
warning!!! the headFlag is not 0x55aa, headFlag:cdac
warning!!! the headFlag is not 0x55aa, headFlag:45ac
duration = 7ms
```

- DAQ框架中实际40个模块单张图像像素数据的模型处理延迟平均为7ms

当前计划



- 对比模型在DAQ框架中推理压缩的图像输出以测试图像数据处理的实际效果
- 研究测试不同硬件平台下模型推理压缩过程在DAQ中的实际性能表现

THANK YOU !