YOLOv9 QAT for TensorRT Detection / Segmentation

README

This repository contains an implementation of YOLOv9 with Quantization-Aware Training (QAT), specifically designed for deployment on platforms utilizing TensorRT for hardware-accelerated inference.
This implementation aims to provide an efficient, low-latency version of YOLOv9 for real-time detection applications.
If you do not intend to deploy your model using TensorRT, it is recommended not to proceed with this implementation.

- The files in this repository represent a patch that adds QAT functionality to the original YOLOv9 repository.
- This patch is intended to be applied to the main YOLOv9 repository to incorporate the ability to train with QAT.
- The implementation is optimized to work efficiently with TensorRT, an inference library that leverages hardware acceleration to enhance inference performance.
- Users interested in implementing object detection using YOLOv9 with QAT on TensorRT platforms can benefit from this repository as it provides a ready-to-use solution.

We use TensorRT's pytorch quntization tool to finetune training QAT yolov9 from the pre-trained weight, then export the model to onnx and deploy it with TensorRT. The accuray and performance can be found in below table.

For those who are not familiar with QAT, I highly recommend watching this video:
Quantization explained with PyTorch - Post-Training Quantization, Quantization-Aware Training

Getting started (工欲善其事，必先利其器)

For getting started, needs some steps.

Git Download

Git

- Settings：git_command

Download Archive

yolov9

```
git clone https://github.com/WongKinYiu/yolov9.git
```

```
git status
git pull
```

yolov9-qat

```
git clone https://github.com/levipereira/yolov9-qat.git
```

```
git status
```

```
git pull
```

CUDA / cudnn Info

Version	Python version	Compiler	Build tools	cuDNN	CUDA
tensorflow-2.14.0	3.9-3.11	Clang 16.0.0	Bazel 6.1.0	8.7	11.8
tensorflow-2.13.0	3.8-3.11	Clang 16.0.0	Bazel 5.3.0	8.6	11.8
tensorflow-2.12.0	3.8-3.11	GCC 9.3.1	Bazel 5.3.0	8.6	11.8
tensorflow-2.11.0	3.7-3.10	GCC 9.3.1	Bazel 5.3.0	8.1	11.2
tensorflow-2.10.0	3.7-3.10	GCC 9.3.1	Bazel 5.1.1	8.1	11.2
tensorflow-2.9.0	3.7-3.10	GCC 9.3.1	Bazel 5.0.0	8.1	11.2
tensorflow-2.8.0	3.7-3.10	GCC 7.3.1	Bazel 4.2.1	8.1	11.2
tensorflow-2.7.0	3.7-3.9	GCC 7.3.1	Bazel 3.7.2	8.1	11.2
tensorflow-2.6.0	3.6-3.9	GCC 7.3.1	Bazel 3.7.2	8.1	11.2
tensorflow-2.5.0	3.6-3.9	GCC 7.3.1	Bazel 3.7.2	8.1	11.2
tensorflow-2.4.0	3.6-3.8	GCC 7.3.1	Bazel 3.1.0	8.0	11.0
tensorflow-2.3.0	3.5-3.8	GCC 7.3.1	Bazel 3.1.0	7.6	10.1
tensorflow-2.2.0	3.5-3.8	GCC 7.3.1	Bazel 2.0.0	7.6	10.1
tensorflow-2.1.0	2.7, 3.3-3.7	GCC 7.3.1	Bazel 0.27.1	7.6	10.1
tensorflow-2.0.0	2.7, 3.3-3.7	GCC 7.3.1	Bazel 0.26.1	7.4	10.0

CUDA cudnn pytorch TensorRT

Implementation Environment

Windows11 + WSL_Ubuntu-20.04(LTS)

Windows11 / WSL_Ubuntu-20.04(LTS)

Windows11

- Environment
  - python 3.10.16
  - CUDA 12.1
  - cudnn 8.8
  - pytorch

```
git clone https://github.com/levipereira/yolov9-qat.git
```

  - TensorRT 10.1.0

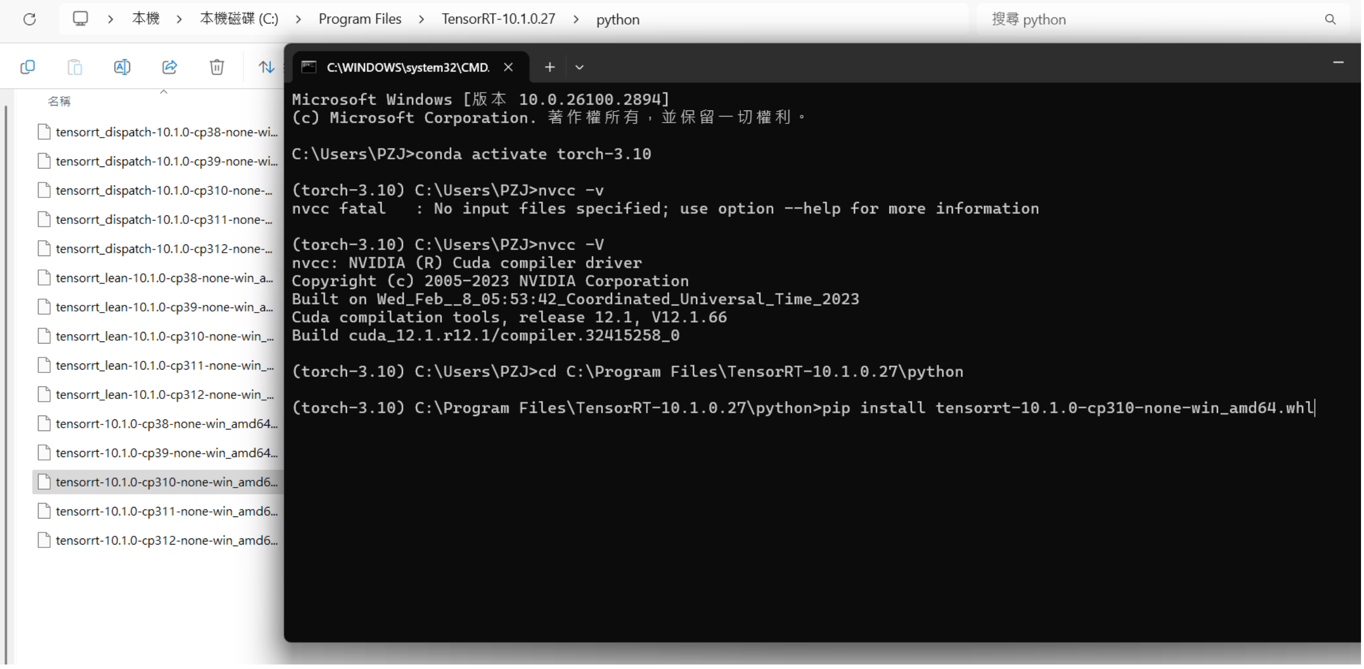
TensorRT Installation Guide

1. Extract the downloaded files.

2. Set the necessary environment variables in the installation directory.

3. Add the `bin` and `lib` folders to the `PATH`.

4. Python_Building


本機 > 本機磁碟 (C:) > Program Files > TensorRT-10.1.0.27 > python
搜尋 python
名稱
C:\WINDOWS\system32\CMD.

tensorrt_dispatch-10.1.0-cp38-none-wi...
tensorrt_dispatch-10.1.0-cp39-none-wi...
tensorrt_dispatch-10.1.0-cp310-none-...
tensorrt_dispatch-10.1.0-cp311-none-...
tensorrt_dispatch-10.1.0-cp312-none-...
tensorrt_lean-10.1.0-cp38-none-win_a...
tensorrt_lean-10.1.0-cp39-none-win_a...
tensorrt_lean-10.1.0-cp310-none-win_...
tensorrt_lean-10.1.0-cp311-none-win_...
tensorrt_lean-10.1.0-cp312-none-win_...
tensorrt-10.1.0-cp38-none-win_amd64...
tensorrt-10.1.0-cp39-none-win_amd64...
tensorrt-10.1.0-cp310-none-win_amd6...
tensorrt-10.1.0-cp311-none-win_amd6...
tensorrt-10.1.0-cp312-none-win_amd6...

```
Microsoft Windows [版本 10.0.26100.2894]
(c) Microsoft Corporation. 著作權所有，並保留一切權利。

C:\Users\PZJ>conda activate torch-3.10

(torch-3.10) C:\Users\PZJ>nvcc -v
nvcc fatal   : No input files specified; use option --help for more information

(torch-3.10) C:\Users\PZJ>nvcc -V
nvcc: NVIDIA (R) Cuda compiler driver
Copyright (c) 2005-2023 NVIDIA Corporation
Built on Wed_Feb__8_05:53:42_Coordinated_Universal_Time_2023
Cuda compilation tools, release 12.1, V12.1.66
Build cuda_12.1.r12.1/compiler.32415258_0

(torch-3.10) C:\Users\PZJ>cd C:\Program Files\TensorRT-10.1.0.27\python

(torch-3.10) C:\Program Files\TensorRT-10.1.0.27\python>pip install tensorrt-10.1.0-cp310-none-win_amd64.whl
```