

# YOLOv11Model conversion + tensorrt acceleration

YOLOv11Model conversion + tensorrt acceleration

1. Model training
  2. Model conversion
- References

After completing the tutorial content of dataset annotation, we can use the motherboard to start training the model.

This tutorial only introduces the model training and conversion of CLI. You can refer to the official website to modify the Python case

## 1. Model training

Use CLI command to train the model directly: copy the yolo11n.pt file to the directory where the configuration file is located, and then open the terminal in the directory where the configuration file is located:

```
cd /home/jetson/ultralytics/ultralytics/data/yahboom_data/orange_data
```

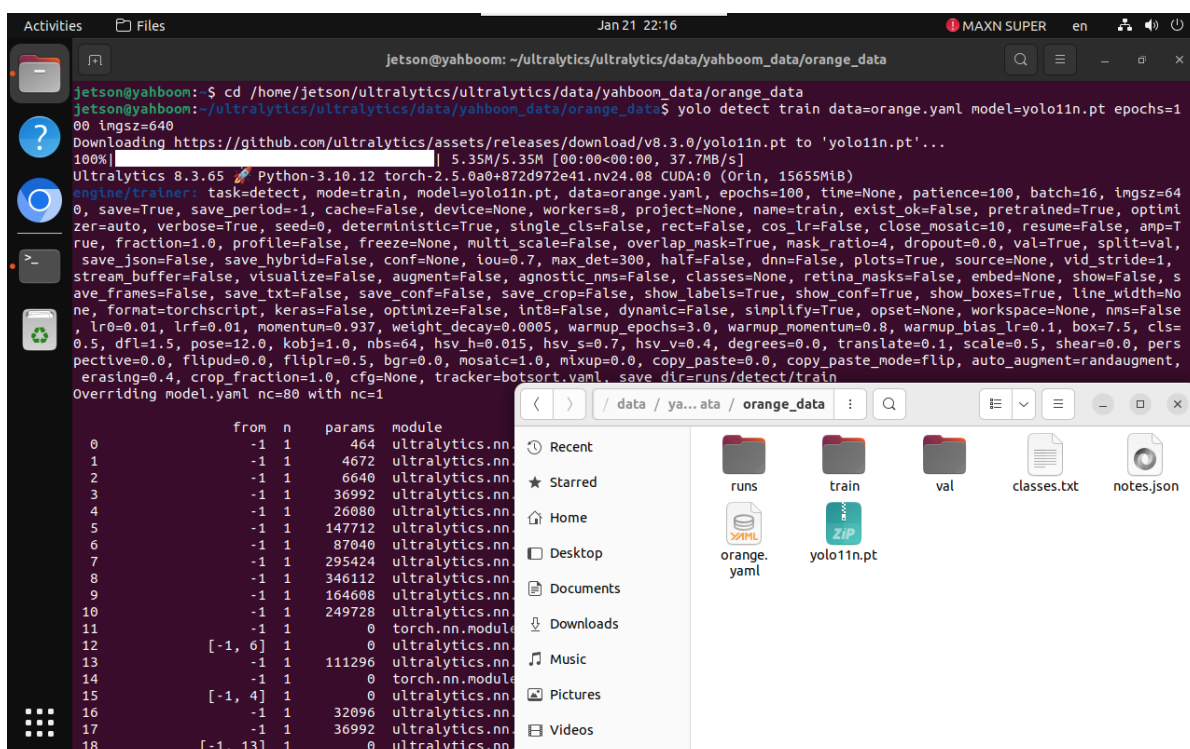
```
yolo detect train data=orange.yaml model=yolo11n.pt epochs=100 imgsz=640
```

**data** : Dataset configuration file

**model** : Pre-trained model file

**epochs** : Number of training rounds

**imgsz** : Input specified image size



```

Activities Terminal Jan 21 22:16 MAXN SUPER en
jetson@yahboom: ~/ultralytics/ultralytics/data/yahboom_data/orange_data

optimizer: 'optimizer=auto' found, ignoring 'lr=0.01' and 'momentum=0.937' and determining best 'optimizer', 'lr' and 'momentum' automatically...
optimizer: AdamW(lr=0.002, momentum=0.9) with parameter groups 81 weight(decay=0.0), 88 weight(decay=0.0005), 87 bias(decay=0.0)
Image sizes 640 train, 640 val
Using 8 dataloader workers
Logging results to runs/detect/train
Starting training for 100 epochs...

Epoch GPU_mem box_loss cls_loss dfl_loss Instances Size
1/100 2.35G 0.5504 2.681 0.983 19 640: 100%| 10/10 [00:07:00:00, 1.39it/s]
Class Images Instances Box(P R mAP50 mAP50-95): 100%| 1/1 [00:01:00:00, 1.44s/
all 20 20 0.00333 1 0.995 0.903

Epoch GPU_mem box_loss cls_loss dfl_loss Instances Size
2/100 2.36G 0.5976 1.759 0.9723 19 640: 100%| 10/10 [00:05:00:00, 1.97it/s]
Class Images Instances Box(P R mAP50 mAP50-95): 100%| 1/1 [00:00:00:00, 3.47it
all 20 20 0.00333 1 0.995 0.901

Epoch GPU_mem box_loss cls_loss dfl_loss Instances Size
3/100 2.38G 0.6172 1.267 1.003 20 640: 100%| 10/10 [00:04:00:00, 2.13it/s]
Class Images Instances Box(P R mAP50 mAP50-95): 100%| 1/1 [00:00:00:00, 3.34it
all 20 20 0.00333 1 0.995 0.864

Epoch GPU_mem box_loss cls_loss dfl_loss Instances Size
4/100 2.38G 0.5826 1.11 0.9764 18 640: 100%| 10/10 [00:04:00:00, 2.08it/s]
Class Images Instances Box(P R mAP50 mAP50-95): 100%| 1/1 [00:00:00:00, 3.36it
all 20 20 0.938 0.76 0.892 0.82

Epoch GPU_mem box_loss cls_loss dfl_loss Instances Size
5/100 2.38G 0.5307 1.014 0.9293 14 640: 100%| 10/10 [00:04:00:00, 2.08it/s]
Class Images Instances Box(P R mAP50 mAP50-95): 100%| 1/1 [00:00:00:00, 3.31it
all 20 20 1 0.76 0.993 0.895

Epoch GPU_mem box_loss cls_loss dfl_loss Instances Size
6/100 2.39G 0.5949 0.9854 0.956 18 640: 100%| 10/10 [00:04:00:00, 2.15it/s]
Class Images Instances Box(P R mAP50 mAP50-95): 100%| 1/1 [00:00:00:00, 3.62it
all 20 20 1 0.839 0.928 0.832

Epoch GPU_mem box_loss cls_loss dfl_loss Instances Size
7/100 2.34G 0.5066 0.8593 0.8889 30 640: 50%| 5/10 [00:02:00:02, 2.13it/s]

```

```

Activities Terminal Jan 22 09:09 MAXN SUPER en
jetson@yahboom: ~/ultralytics/ultralytics/data/yahboom_data/orange_data

all 20 20 0.997 1 0.995 0.945

Epoch GPU_mem box_loss cls_loss dfl_loss Instances Size
96/100 2.38G 0.2816 0.222 0.8046 9 640: 100%| 10/10 [00:04:00:00, 2.28it/s]
Class Images Instances Box(P R mAP50 mAP50-95): 100%| 1/1 [00:00:00:00, 3.65it
all 20 20 0.997 1 0.995 0.929

Epoch GPU_mem box_loss cls_loss dfl_loss Instances Size
97/100 2.37G 0.2746 0.2167 0.7994 9 640: 100%| 10/10 [00:04:00:00, 2.32it/s]
Class Images Instances Box(P R mAP50 mAP50-95): 100%| 1/1 [00:00:00:00, 3.69it
all 20 20 0.997 1 0.995 0.929

Epoch GPU_mem box_loss cls_loss dfl_loss Instances Size
98/100 2.38G 0.2702 0.2085 0.8157 9 640: 100%| 10/10 [00:04:00:00, 2.30it/s]
Class Images Instances Box(P R mAP50 mAP50-95): 100%| 1/1 [00:00:00:00, 3.74it
all 20 20 0.997 1 0.995 0.934

Epoch GPU_mem box_loss cls_loss dfl_loss Instances Size
99/100 2.37G 0.262 0.2092 0.7864 9 640: 100%| 10/10 [00:04:00:00, 2.32it/s]
Class Images Instances Box(P R mAP50 mAP50-95): 100%| 1/1 [00:00:00:00, 3.79it
all 20 20 0.997 1 0.995 0.937

Epoch GPU_mem box_loss cls_loss dfl_loss Instances Size
100/100 2.38G 0.258 0.2032 0.8043 9 640: 100%| 10/10 [00:04:00:00, 2.30it/s]
Class Images Instances Box(P R mAP50 mAP50-95): 100%| 1/1 [00:00:00:00, 3.81it
all 20 20 0.997 1 0.995 0.939

100 epochs completed in 0.154 hours.
Optimizer stripped from runs/detect/train2/weights/last.pt, 5.5MB
Optimizer stripped from runs/detect/train2/weights/best.pt, 5.5MB

Validating runs/detect/train2/weights/best.pt...
Ultralytics 8.3.65 Python-3.10.12 torch-2.5.0a0+872d972e41.nv24.08 CUDA:0 (Orin, 15655MiB)
YOLO11n summary (fused): 238 layers, 2,582,347 parameters, 0 gradients, 6.3 GFLOPs
Class Images Instances Box(P R mAP50 mAP50-95): 100%| 1/1 [00:00:00:00, 3.62it
all 20 20 0.997 1 0.995 0.952

Speed: 0.3ms preprocess, 6.1ms inference, 0.0ms loss, 1.4ms postprocess per image
Results saved to runs/detect/train2
Learn more at https://docs.ultralytics.com/modes/train
jetson@yahboom: ~/ultralytics/ultralytics/data/yahboom_data/orange_data$

```

## 2. Model conversion

The final model will be generated in the runs folder: generally select the best.pt file for use

```
/home/jetson/ultralytics/ultralytics/data/yahboom_data/orange_data/runs/detect/train/weights
```

Convert the PyTorch format model to TensorRT:

```
cd
/home/jetson/ultralytics/ultralytics/data/yahboom_data/orange_data/runs/detect/train/weights
```

```
yolo export model=best.pt format=engine
```

```
Activities Terminal Jan 22 09:42 MAXN SUPER en
jetson@yahboom: ~/ultralytics/ultralytics/data/yahboom_data/orange_data/runs/detect/train/weights
jetson@yahboom:~/ultralytics/ultralytics/data/yahboom_data/orange_data/runs/detect/train/weights$ yolo export model=best.pt format=engine
WARNING ⚠️TensorRT requires GPU export, automatically assigning device=0
Ultralytics 8.3.65 Python-3.10.12 torch-2.5.0a0+872d972e41.nv24.00 CUDA:0 (Orin, 15655MiB)
YOLO11n summary (fused): 238 layers, 2,502,347 parameters, 0 gradients, 6.3 GFLOPs

PyTorch: starting from 'best.pt' with input shape (1, 3, 640, 640) BCHW and output shape(s) (1, 5, 8400) (5.2 MB)

ONNX: starting export with onnx 1.17.0 opset 19...
ONNX: slimming with onnxslim 0.1.47...
ONNX: export success ✅ 1.9s, saved as 'best.onnx' (10.1 MB)

TensorRT: starting export with TensorRT 10.7.0...
[01/22/2025-09:10:31] [TRT] [I] [MemUsageChange] Init CUDA: CPU -1, GPU +0, now: CPU 727, GPU 4907 (MiB)
[01/22/2025-09:10:33] [TRT] [I] [MemUsageChange] Init builder kernel library: CPU +980, GPU +802, now: CPU 1719, GPU 5739 (MiB)
[01/22/2025-09:10:33] [TRT] [I] -----
[01/22/2025-09:10:33] [TRT] [I] Input filename: best.onnx
[01/22/2025-09:10:33] [TRT] [I] ONNX IR version: 0.0.9
[01/22/2025-09:10:33] [TRT] [I] Opset version: 19
[01/22/2025-09:10:33] [TRT] [I] Producer name: pytorch
[01/22/2025-09:10:33] [TRT] [I] Producer version: 2.5.0
[01/22/2025-09:10:33] [TRT] [I] Domain:
[01/22/2025-09:10:33] [TRT] [I] Model version: 0
[01/22/2025-09:10:33] [TRT] [I] Doc string:
[01/22/2025-09:10:33] [TRT] [I] -----
TensorRT: input "images" with shape(1, 3, 640, 640) DataType.FLOAT
TensorRT: output "output0" with shape(1, 5, 8400) DataType.FLOAT
TensorRT: building FP32 engine as best.engine
[01/22/2025-09:10:34] [TRT] [I] Local timing cache in use. Profiling results in this builder pass will not be stored.
[01/22/2025-09:11:33] [TRT] [I] Compiler backend is used during engine build.
[01/22/2025-09:12:37] [TRT] [I] Detected 1 inputs and 1 output network tensors.
[01/22/2025-09:12:38] [TRT] [I] Total Host Persistent Memory: 557632 bytes
[01/22/2025-09:12:38] [TRT] [I] Total Device Persistent Memory: 0 bytes
[01/22/2025-09:12:38] [TRT] [I] Max Scratch Memory: 2764800 bytes
[01/22/2025-09:12:38] [TRT] [I] [BlockAssignment] Started assigning block shifts. This will take 234 steps to complete.
[01/22/2025-09:12:38] [TRT] [I] [BlockAssignment] Algorithm ShiftNTopDown took 18.7229ms to assign 10 blocks to 234 nodes requiring 19046912 bytes.
[01/22/2025-09:12:38] [TRT] [I] Total Activation Memory: 19046400 bytes
[01/22/2025-09:12:38] [TRT] [I] Total Weights Memory: 10456132 bytes
[01/22/2025-09:12:38] [TRT] [I] Compiler backend is used during engine execution.
```

```
Activities Terminal Jan 22 09:42 MAXN SUPER en
jetson@yahboom: ~/ultralytics/ultralytics/data/yahboom_data/orange_data/runs/detect/train/weights
[01/22/2025-09:10:31] [TRT] [I] [MemUsageChange] Init CUDA: CPU -1, GPU +0, now: CPU 727, GPU 4907 (MiB)
[01/22/2025-09:10:33] [TRT] [I] [MemUsageChange] Init builder kernel library: CPU +980, GPU +802, now: CPU 1719, GPU 5739 (MiB)
[01/22/2025-09:10:33] [TRT] [I] -----
[01/22/2025-09:10:33] [TRT] [I] Input filename: best.onnx
[01/22/2025-09:10:33] [TRT] [I] ONNX IR version: 0.0.9
[01/22/2025-09:10:33] [TRT] [I] Opset version: 19
[01/22/2025-09:10:33] [TRT] [I] Producer name: pytorch
[01/22/2025-09:10:33] [TRT] [I] Producer version: 2.5.0
[01/22/2025-09:10:33] [TRT] [I] Domain:
[01/22/2025-09:10:33] [TRT] [I] Model version: 0
[01/22/2025-09:10:33] [TRT] [I] Doc string:
[01/22/2025-09:10:33] [TRT] [I] -----
TensorRT: input "images" with shape(1, 3, 640, 640) DataType.FLOAT
TensorRT: output "output0" with shape(1, 5, 8400) DataType.FLOAT
TensorRT: building FP32 engine as best.engine
[01/22/2025-09:10:34] [TRT] [I] Local timing cache in use. Profiling results in this builder pass will not be stored.
[01/22/2025-09:11:33] [TRT] [I] Compiler backend is used during engine build.
[01/22/2025-09:12:37] [TRT] [I] Detected 1 inputs and 1 output network tensors.
[01/22/2025-09:12:38] [TRT] [I] Total Host Persistent Memory: 557632 bytes
[01/22/2025-09:12:38] [TRT] [I] Total Device Persistent Memory: 0 bytes
[01/22/2025-09:12:38] [TRT] [I] Max Scratch Memory: 2764800 bytes
[01/22/2025-09:12:38] [TRT] [I] [BlockAssignment] Started assigning block shifts. This will take 234 steps to complete.
[01/22/2025-09:12:38] [TRT] [I] [BlockAssignment] Algorithm ShiftNTopDown took 18.7229ms to assign 10 blocks to 234 nodes requiring 19046912 bytes.
[01/22/2025-09:12:38] [TRT] [I] Total Activation Memory: 19046400 bytes
[01/22/2025-09:12:38] [TRT] [I] Total Weights Memory: 10456132 bytes
[01/22/2025-09:12:38] [TRT] [I] Compiler backend is used during engine execution.
[01/22/2025-09:12:38] [TRT] [I] Engine generation completed in 124.854 seconds.
[01/22/2025-09:12:38] [TRT] [I] [MemUsageStats] Peak memory usage of TRT CPU/GPU memory allocators: CPU 1 MiB, GPU 132 MiB
TensorRT: export success ✅ 129.4s, saved as 'best.engine' (11.9 MB)

Export complete (130.4s)
Results saved to /home/jetson/ultralytics/ultralytics/data/yahboom_data/orange_data/runs/detect/train/weights
Predict: yolo predict task=detect model=best.engine imgs=640
Validate: yolo val task=detect model=best.engine imgs=640 data=orange.yaml
Visualize: https://netron.app
💡 Learn more at https://docs.ultralytics.com/modes/export
jetson@yahboom:~/ultralytics/ultralytics/data/yahboom_data/orange_data/runs/detect/train/weights$ ls
best.engine best.onnx best.pt last.pt
jetson@yahboom:~/ultralytics/ultralytics/data/yahboom_data/orange_data/runs/detect/train/weights$
```

## References

<https://docs.ultralytics.com/modes/train/>