

# Vitis AI Lab 1 & 2

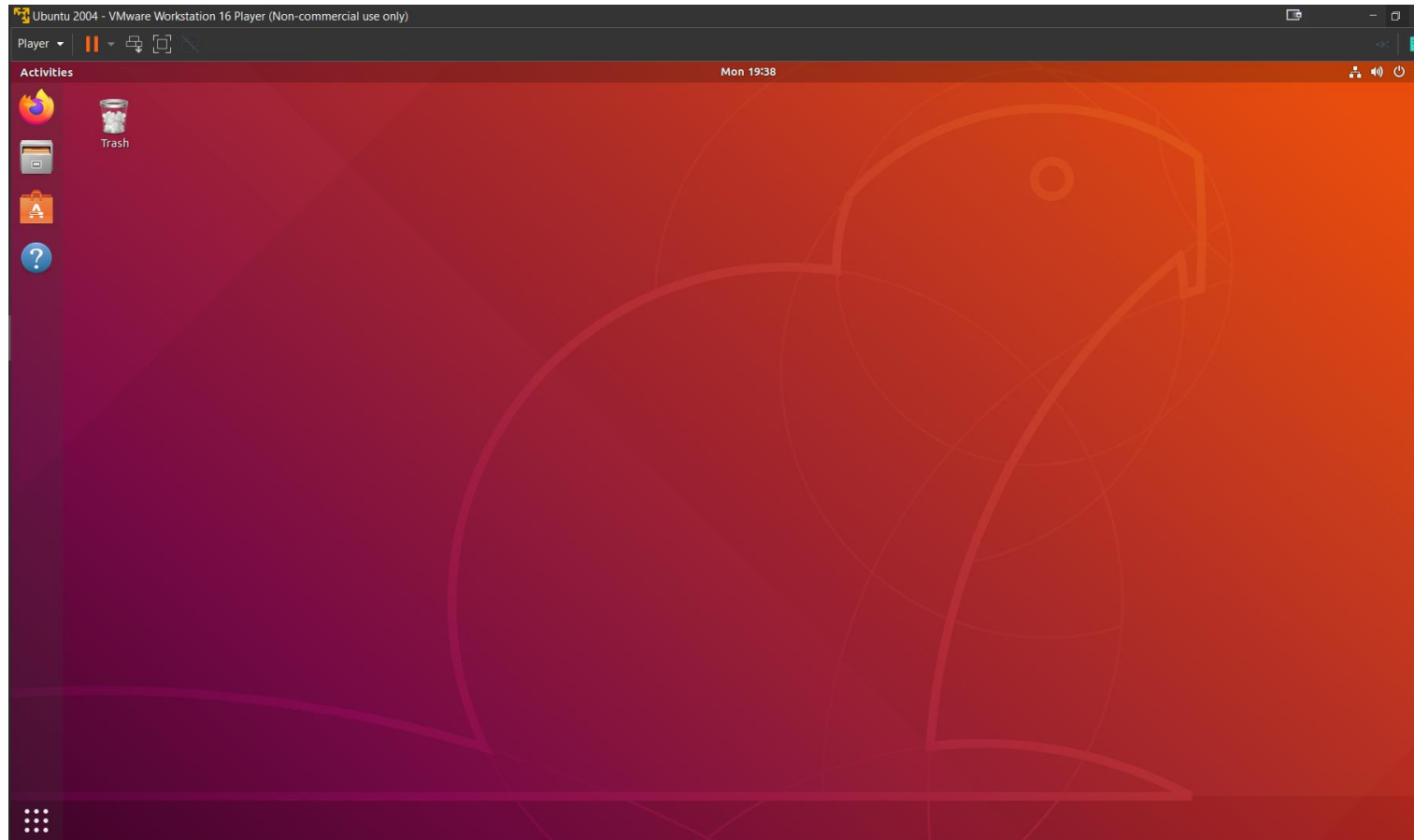
# Agenda

- Environment Setup
- Lab 1: AI Quantizer and AI Compiler – Caffe
- Lab 2: AI Quantizer and AI Compiler – TensorFlow2 and PyTorch

# Environment Setup

# Environment Setup

- Ubuntu 20.04 on VMware
  1. Install VMware
  2. Download the iso file of Ubuntu 20.04



# Environment Setup

- Docker

1. Install Docker

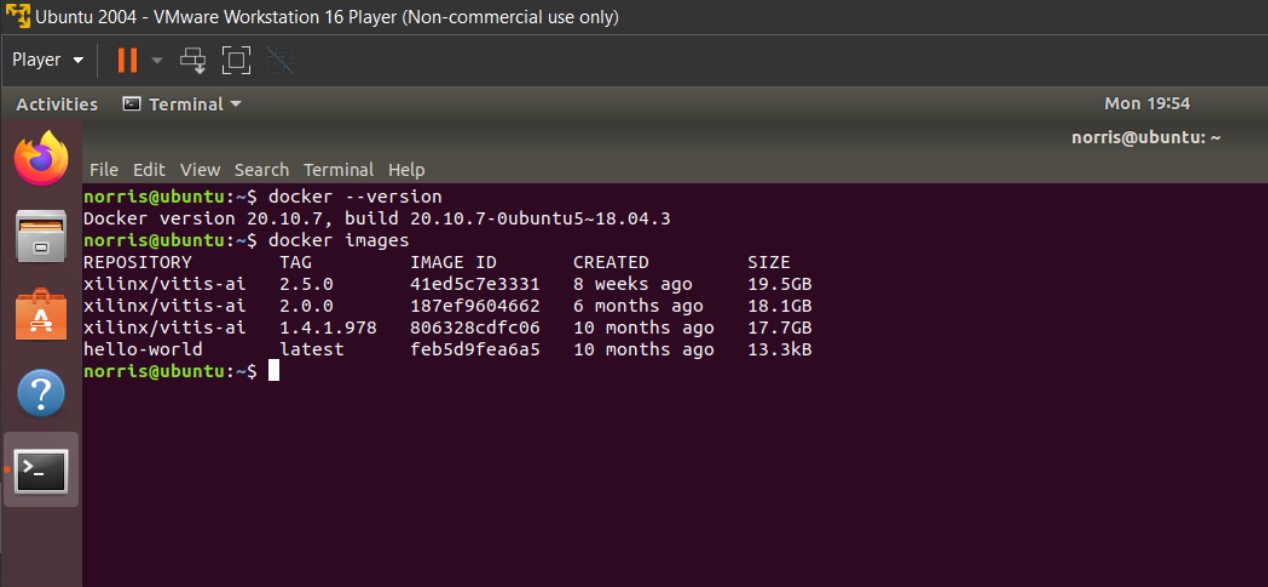
```
sudo apt-get install docker.io
```

2. Check Docker installation

```
sudo docker --version
```

3. Remove typing 'sudo' before docker

```
sudo chmod 666 /var/run/docker.sock
```



The screenshot shows a terminal window titled 'Ubuntu 2004 - VMware Workstation 16 Player (Non-commercial use only)'. The terminal output shows the Docker version and a list of installed images.

```
norris@ubuntu:~$ docker --version
Docker version 20.10.7, build 20.10.7-0ubuntu5~18.04.3
norris@ubuntu:~$ docker images
```

REPOSITORY	TAG	IMAGE ID	CREATED	SIZE
xilinx/vitis-ai	2.5.0	41ed5c7e3331	8 weeks ago	19.5GB
xilinx/vitis-ai	2.0.0	187ef9604662	6 months ago	18.1GB
xilinx/vitis-ai	1.4.1.978	806328cdfc06	10 months ago	17.7GB
hello-world	latest	feb5d9fea6a5	10 months ago	13.3kB

# Environment Setup

- Vitis AI 2.0
  1. Download Vitis AI v2.0 from github  
`git clone https://github.com/Xilinx/Vitis-AI.git -b v2.0`
  2. Download Vitis AI Docker Image  
`docker pull xilinx/vitis-ai:2.0.0`
  3. Activate Vitis AI Docker Environment  
`./docker_run.sh 2.0.0`

```
=====
VITIS-AI
=====

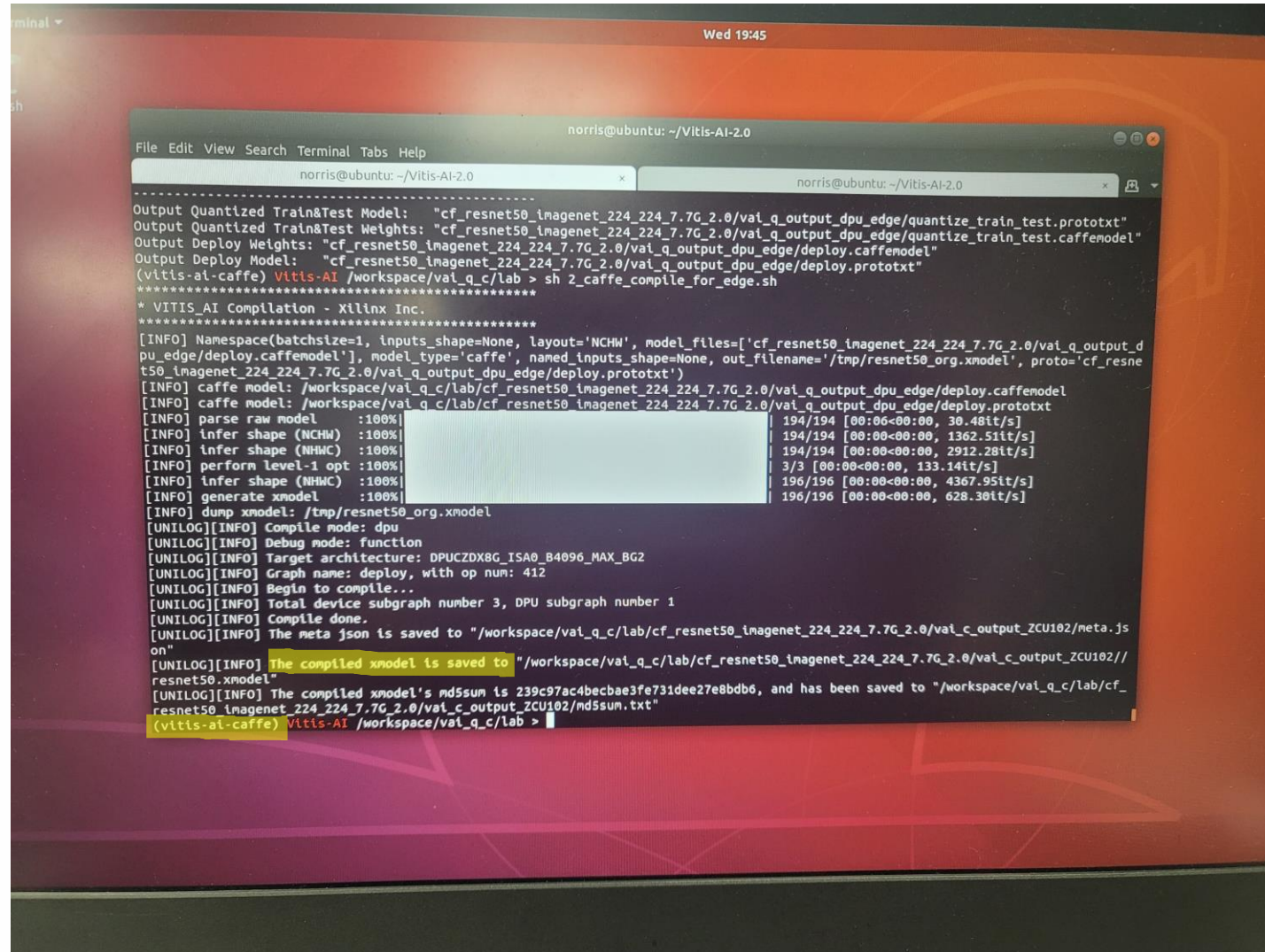
Docker Image Version: 2.0.0.1103 (CPU)
Vitis AI Git Hash: 06d7cbb
Build Date: 2022-01-12

For TensorFlow 1.15 Workflows do:
    conda activate vitis-ai-tensorflow
For Caffe Workflows do:
    conda activate vitis-ai-caffe
For PyTorch Workflows do:
    conda activate vitis-ai-pytorch
For TensorFlow 2.6 Workflows do:
    conda activate vitis-ai-tensorflow2
Vitis-AI /workspace > |
```

# Lab 1: AI Quantizer and AI Compiler – Caffe

# AI Quantizer and AI Compiler – Caffe

- Result - Caffe



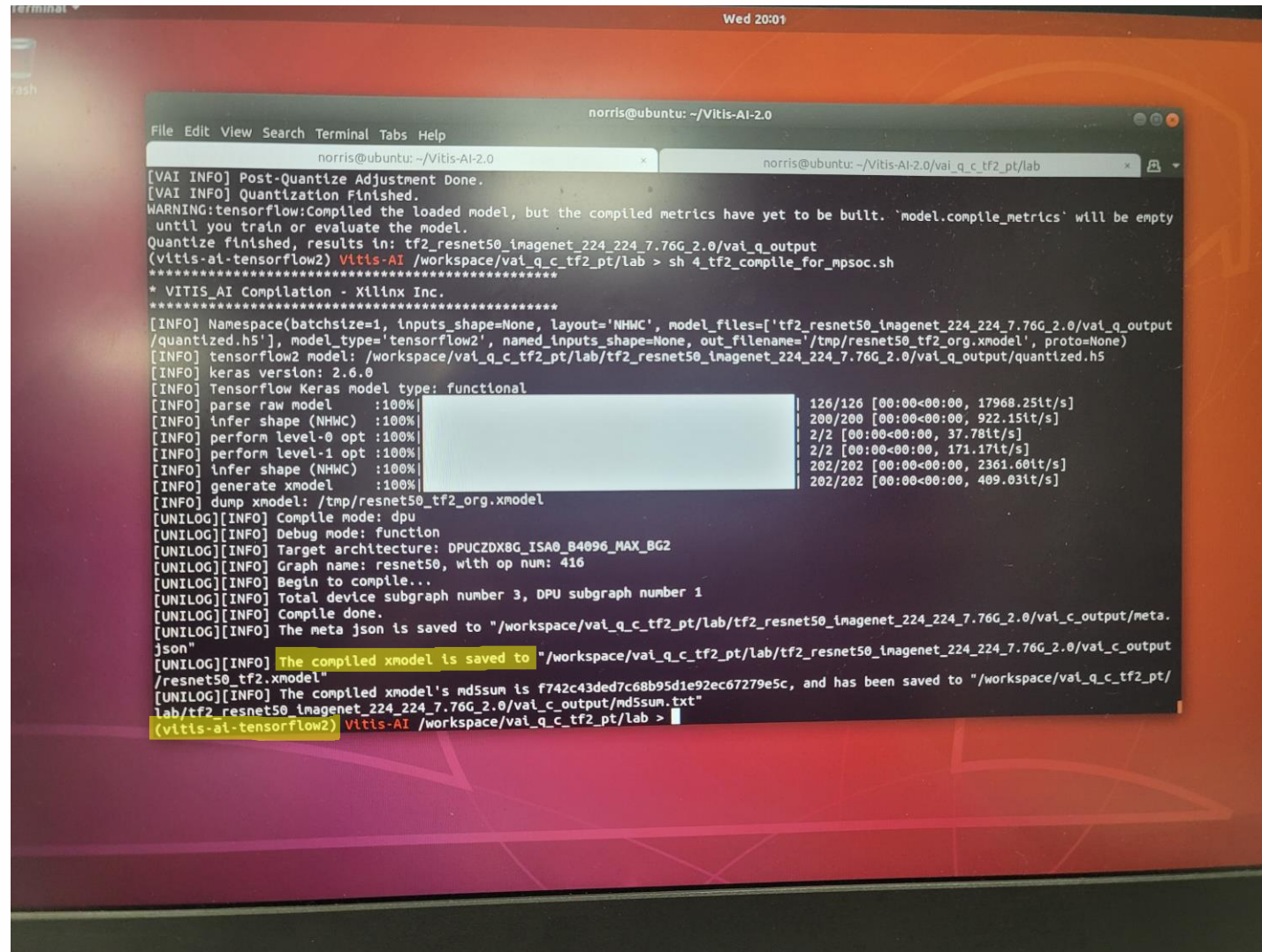
```
norris@ubuntu: ~/Vitis-AI-2.0
File Edit View Search Terminal Tabs Help
norris@ubuntu: ~/Vitis-AI-2.0
Output Quantized Train&Test Model: "cf_resnet50_imagenet_224_224_7.7G_2.0/vai_q_output_dpu_edge/quantize_train_test.prototxt"
Output Quantized Train&Test Weights: "cf_resnet50_imagenet_224_224_7.7G_2.0/vai_q_output_dpu_edge/quantize_train_test.caffemodel"
Output Deploy Weights: "cf_resnet50_imagenet_224_224_7.7G_2.0/vai_q_output_dpu_edge/deploy.caffemodel"
Output Deploy Model: "cf_resnet50_imagenet_224_224_7.7G_2.0/vai_q_output_dpu_edge/deploy.prototxt"
(vitis-ai-caffe) Vitis-AI /workspace/vai_q_c/lab > sh 2_caffe_compile_for_edge.sh
*****
* VITIS_AI Compilation - Xilinx Inc.
*****
[INFO] Namespace(batchsize=1, inputs_shape=None, layout='NCHW', model_files=['cf_resnet50_imagenet_224_224_7.7G_2.0/vai_q_output_dpu_edge/deploy.caffemodel'], model_type='caffe', named_inputs_shape=None, out_filename='/tmp/resnet50_org.xmodel', proto='cf_resnet50_imagenet_224_224_7.7G_2.0/vai_q_output_dpu_edge/deploy.prototxt')
[INFO] caffe model: /workspace/vai_q_c/lab/cf_resnet50_imagenet_224_224_7.7G_2.0/vai_q_output_dpu_edge/deploy.caffemodel
[INFO] caffe model: /workspace/vai_q_c/lab/cf_resnet50_imagenet_224_224_7.7G_2.0/vai_q_output_dpu_edge/deploy.prototxt
[INFO] parse raw model :100%|
[INFO] infer shape (NCHW) :100%|
[INFO] infer shape (NHWC) :100%|
[INFO] perform level-1 opt :100%|
[INFO] infer shape (NHWC) :100%|
[INFO] generate xmodel :100%|
[INFO] dump xmodel: /tmp/resnet50_org.xmodel
[UNILog][INFO] Compile mode: dpu
[UNILog][INFO] Debug mode: function
[UNILog][INFO] Target architecture: DPUCZDX8G ISA0_B4096_MAX_BG2
[UNILog][INFO] Graph name: deploy, with op num: 412
[UNILog][INFO] Begin to compile...
[UNILog][INFO] Total device subgraph number 3, DPU subgraph number 1
[UNILog][INFO] Compile done.
[UNILog][INFO] The meta json is saved to "/workspace/vai_q_c/lab/cf_resnet50_imagenet_224_224_7.7G_2.0/vai_c_output_ZCU102/meta.json"
[UNILog][INFO] The compiled xmodel is saved to "/workspace/vai_q_c/lab/cf_resnet50_imagenet_224_224_7.7G_2.0/vai_c_output_ZCU102/resnet50.xmodel"
[UNILog][INFO] The compiled xmodel's md5sum is 239c97ac4becbae3fe731dee27e8bdb6, and has been saved to "/workspace/vai_q_c/lab/cf_resnet50_imagenet_224_224_7.7G_2.0/vai_c_output_ZCU102/md5sum.txt"
(vitis-ai-caffe) Vitis-AI /workspace/vai_q_c/lab >
```



# Lab 2: AI Quantizer and AI Compiler – TensorFlow2 and PyTorch

# AI Quantizer and AI Compiler – TensorFlow2 and PyTorch

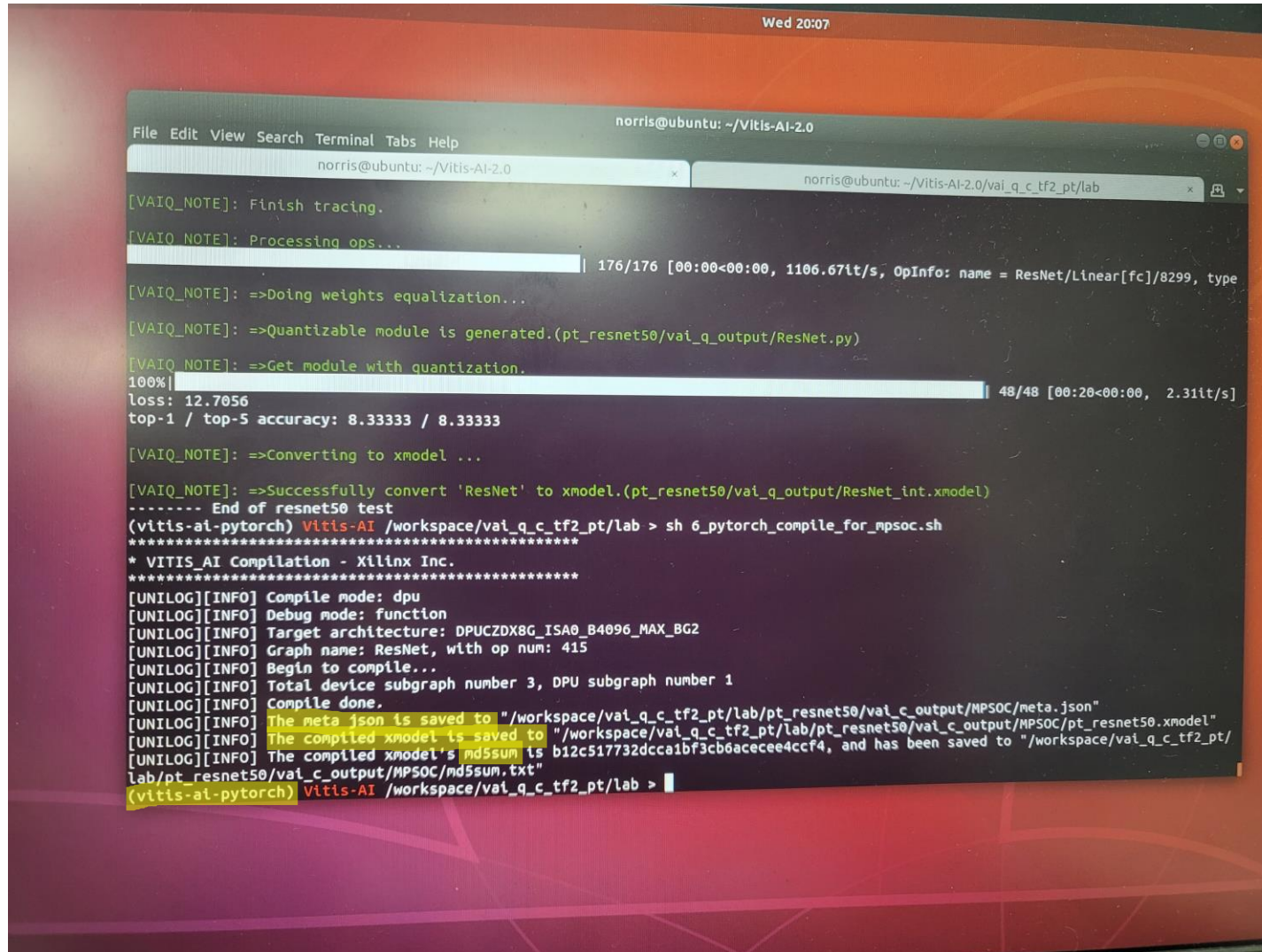
- Result – Tensorflow2



```
norris@ubuntu: ~/Vitis-AI-2.0
[VAI INFO] Post-Quantize Adjustment Done.
[VAI INFO] Quantization Finished.
WARNING:tensorflow:Compiled the loaded model, but the compiled metrics have yet to be built. 'model.compile_metrics' will be empty
until you train or evaluate the model.
Quantize finished, results in: tf2_resnet50_imagenet_224_224_7.76G_2.0/vai_q_output
(vitis-ai-tensorflow2) Vitis-AI /workspace/vai_q_c_tf2_pt/lab > sh 4_tf2_compile_for_mpsoc.sh
*****
* Vitis AI Compilation - Xilinx Inc.
*****
[INFO] Namespace(batchsize=1, inputs_shape=None, layout='NHWC', model_files=['tf2_resnet50_imagenet_224_224_7.76G_2.0/vai_q_output
/quantized.h5'], model_type='tensorflow2', named_inputs_shape=None, out_filename='/tmp/resnet50_tf2_org.xmodel', proto=None)
[INFO] tensorflow2 model: /workspace/vai_q_c_tf2_pt/lab/tf2_resnet50_imagenet_224_224_7.76G_2.0/vai_q_output/quantized.h5
[INFO] keras version: 2.6.0
[INFO] Tensorflow Keras model type: functional
[INFO] parse raw model :100%| 126/126 [00:00<00:00, 17968.25it/s]
[INFO] infer shape (NHWC) :100%| 200/200 [00:00<00:00, 922.15it/s]
[INFO] perform level-0 opt :100%| 2/2 [00:00<00:00, 37.78it/s]
[INFO] perform level-1 opt :100%| 2/2 [00:00<00:00, 171.17it/s]
[INFO] infer shape (NHWC) :100%| 202/202 [00:00<00:00, 2361.60it/s]
[INFO] generate xmodel :100%| 202/202 [00:00<00:00, 409.03it/s]
[INFO] dump xmodel: /tmp/resnet50_tf2_org.xmodel
[UNILog][INFO] Compile mode: dpu
[UNILog][INFO] Debug mode: function
[UNILog][INFO] Target architecture: DPUCZDX8G_ISA0_B4096_MAX_BG2
[UNILog][INFO] Graph name: resnet50, with op num: 416
[UNILog][INFO] Begin to compile...
[UNILog][INFO] Total device subgraph number 3, DPU subgraph number 1
[UNILog][INFO] Compile done.
[UNILog][INFO] The meta json is saved to "/workspace/vai_q_c_tf2_pt/lab/tf2_resnet50_imagenet_224_224_7.76G_2.0/vai_c_output/meta.
json"
[UNILog][INFO] The compiled xmodel is saved to "/workspace/vai_q_c_tf2_pt/lab/tf2_resnet50_imagenet_224_224_7.76G_2.0/vai_c_output
/resnet50_tf2.xmodel"
[UNILog][INFO] The compiled xmodel's md5sum is f742c43ded7c68b95d1e92ec67279e5c, and has been saved to "/workspace/vai_q_c_tf2_pt/
lab/tf2_resnet50_imagenet_224_224_7.76G_2.0/vai_c_output/md5sum.txt"
(vitis-ai-tensorflow2) Vitis-AI /workspace/vai_q_c_tf2_pt/lab >
```

# AI Quantizer and AI Compiler – TensorFlow2 and PyTorch

- Result - PyTorch



```
norris@ubuntu: ~/Vitis-AI-2.0
File Edit View Search Terminal Tabs Help
norris@ubuntu: ~/Vitis-AI-2.0
[VAIQ_NOTE]: Finish tracing.
[VAIQ_NOTE]: Processing ops...
| 176/176 [00:00<00:00, 1106.67it/s, OpInfo: name = ResNet/Linear[fc]/8299, type
[VAIQ_NOTE]: =>Doing weights equalization...
[VAIQ_NOTE]: =>Quantizable module is generated.(pt_resnet50/vai_q_output/ResNet.py)
[VAIQ_NOTE]: =>Get module with quantization.
100%|
loss: 12.7056
top-1 / top-5 accuracy: 8.33333 / 8.33333
[VAIQ_NOTE]: =>Converting to xmodel ...
[VAIQ_NOTE]: =>Successfully convert 'ResNet' to xmodel.(pt_resnet50/vai_q_output/ResNet_int.xmodel)
----- End of resnet50 test
(vitis-ai-pytorch) Vitis-AI /workspace/vai_q_c_tf2_pt/lab > sh 6_pytorch_compile_for_mpsoc.sh
*****
* VITIS_AI Compilation - Xilinx Inc.
*****
[UNILog][INFO] Compile mode: dpu
[UNILog][INFO] Debug mode: function
[UNILog][INFO] Target architecture: DPUCZDX8G_ISA0_B4096_MAX_BG2
[UNILog][INFO] Graph name: ResNet, with op num: 415
[UNILog][INFO] Begin to compile...
[UNILog][INFO] Total device subgraph number 3, DPU subgraph number 1
[UNILog][INFO] Compile done.
[UNILog][INFO] The meta json is saved to "/workspace/vai_q_c_tf2_pt/lab/pt_resnet50/vai_q_output/MP5OC/meta.json"
[UNILog][INFO] The compiled xmodel is saved to "/workspace/vai_q_c_tf2_pt/lab/pt_resnet50/vai_q_output/MP5OC/pt_resnet50.xmodel"
[UNILog][INFO] The compiled xmodel's md5sum is b12c517732dcca1bf3cb6acece4ccf4, and has been saved to "/workspace/vai_q_c_tf2_pt/
lab/pt_resnet50/vai_q_output/MP5OC/md5sum.txt"
(vitis-ai-pytorch) Vitis-AI /workspace/vai_q_c_tf2_pt/lab >
```