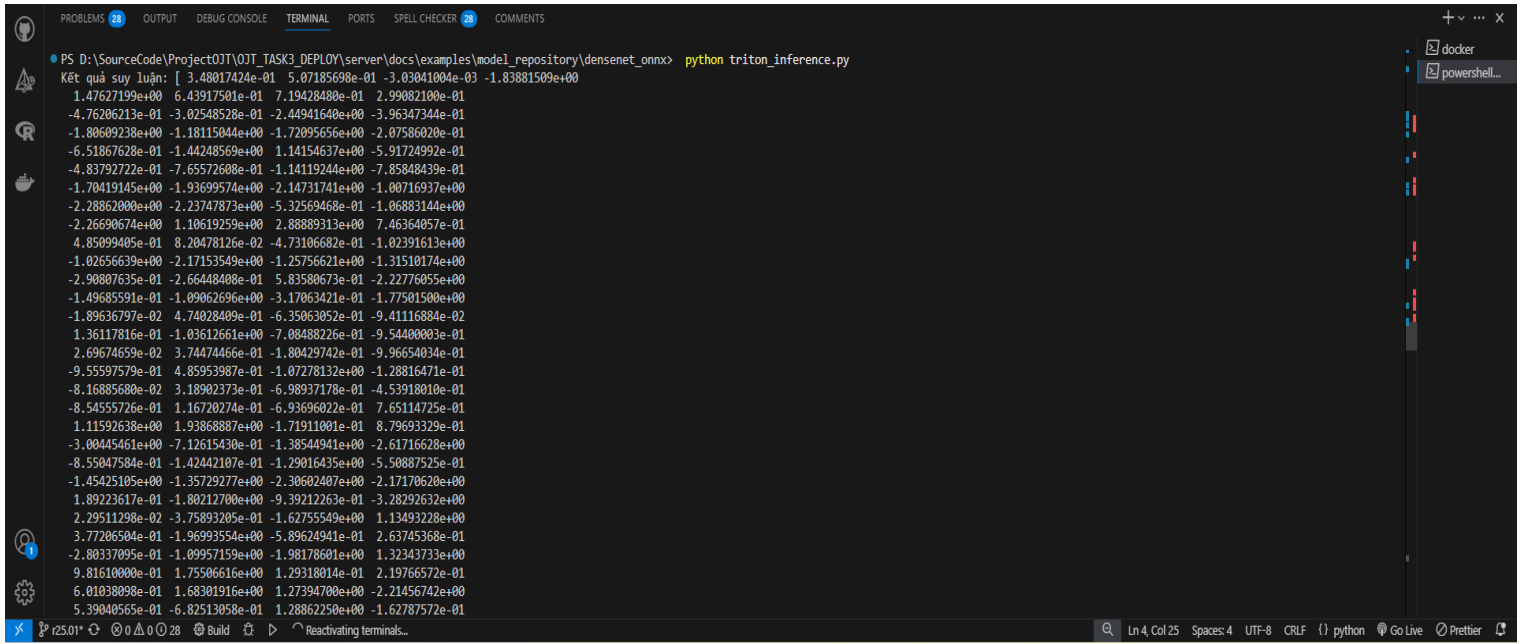


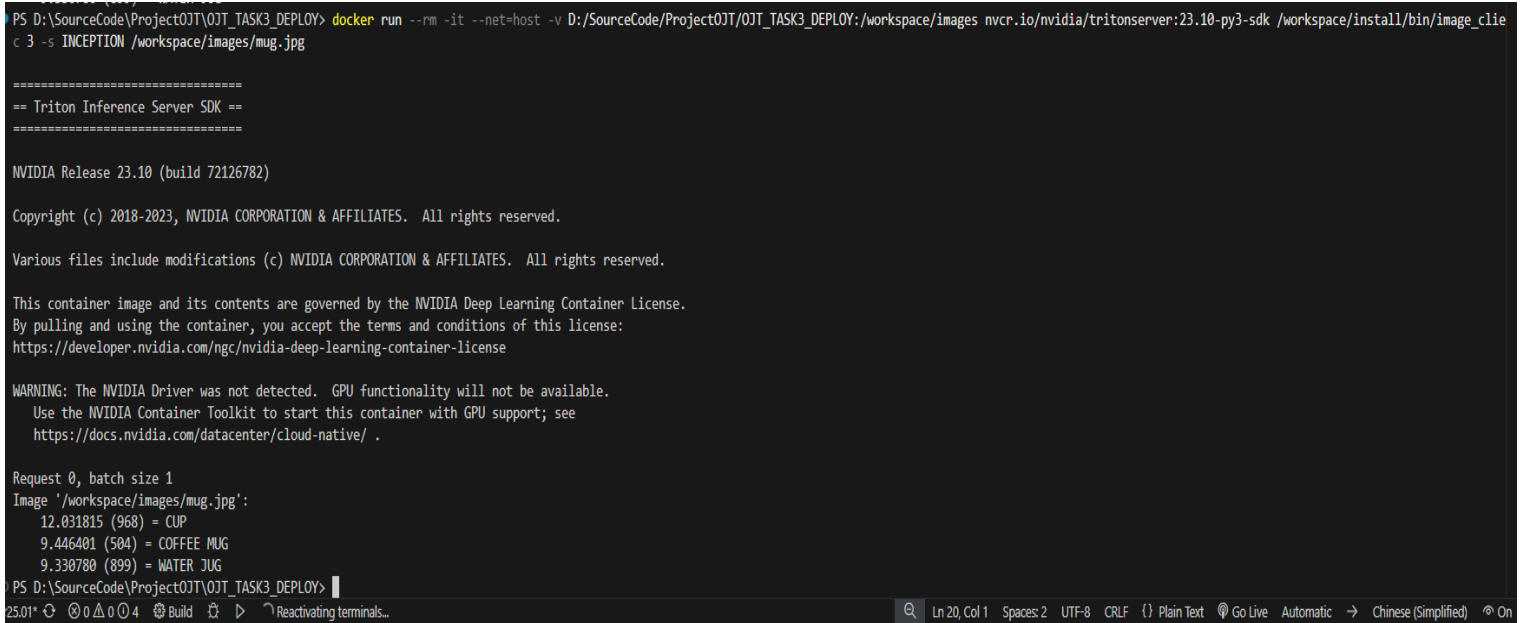
FINAL - RESULT

1 . Result Inference



```
PS D:\SourceCode\ProjectOJT\OJT_TASK3_DEPLOY\server\docs\examples\model_repository\densenet_onnx> python triton_inference.py
K&#228;t quả suy lu&#223;n: [ 3.48017424e-01 5.07185698e-01 -3.03041004e-03 -1.83881509e+00
1.47627199e+00 6.43917501e-01 7.19428480e-01 2.99082100e-01
-4.76206213e-01 -3.02548528e-01 -2.44941640e+00 -3.96347344e-01
-1.80609238e+00 -1.18115044e+00 -1.72095656e+00 -2.07586020e-01
-6.51867628e-01 -1.44248569e+00 1.14154637e+00 -5.91724992e-01
-4.83792722e-01 -7.65572608e-01 -1.14119244e+00 -7.85848439e-01
-1.70419145e+00 -1.93699574e+00 -2.14731741e+00 -1.00716937e+00
-2.28862000e+00 -2.23747873e+00 -5.32569468e-01 -1.06883144e+00
-2.26690674e+00 1.10619259e+00 2.88809313e+00 7.46364057e-01
4.85099405e-01 8.20478126e-02 -4.73106682e-01 -1.02391613e+00
-1.02656639e+00 -2.17153549e+00 -1.25756621e+00 -1.31510174e+00
-2.90807635e-01 -2.66448408e-01 5.83580673e-01 -2.22776055e+00
-1.49685591e-01 -1.09062696e+00 -3.17063421e-01 -1.77501500e+00
-1.89636797e-02 4.74028409e-01 -6.35063052e-01 -9.41116884e-02
1.36117816e-01 -1.03612661e+00 -7.08488226e-01 -9.54400003e-01
2.69674659e-02 3.74474466e-01 -1.80429742e-01 -9.96654034e-01
-9.55597579e-01 4.85953987e-01 -1.07278132e+00 -1.28816471e-01
-8.16885680e-02 3.18902373e-01 -6.98937178e-01 -4.53918010e-01
-8.5455726e-01 1.16720274e-01 -6.93696022e-01 7.65114725e-01
1.11592638e+00 1.93868887e+00 -1.71911001e-01 8.79693329e-01
-3.00445461e+00 -7.12615430e-01 -1.38544941e+00 -2.61716628e+00
-8.55047584e-01 -1.42442107e-01 -1.29016435e+00 -5.50887525e-01
-1.45425105e+00 -1.35729277e+00 -2.30602407e+00 -2.17170620e+00
1.89223617e-01 -1.80212700e+00 -9.39212263e-01 -3.28292632e+00
2.29511298e-02 -3.75893205e-01 -1.62755549e+00 1.13493228e+00
3.77206504e-01 -1.96993554e+00 -5.89624941e-01 2.63745368e-01
-2.80337095e-01 -1.09957159e+00 -1.98178601e+00 1.32343733e+00
9.81610000e-01 1.75506616e+00 1.29318014e-01 2.19766572e-01
6.01038098e-01 1.68301916e+00 1.27394700e+00 -2.21456742e+00
5.39040565e-01 -6.82513058e-01 1.28862250e+00 -1.62787572e-01
```

2: Result:



```
PS D:\SourceCode\ProjectOJT\OJT_TASK3_DEPLOY> docker run --rm -it --net=host -v D:\SourceCode\ProjectOJT\OJT_TASK3_DEPLOY\workspace\images nvcr.io/nvidia/tritonserver:23.10-py3-sdk /workspace/install/bin/image_cli
c 3 -s INCEPTION /workspace/images/mug.jpg

=====
== Triton Inference Server SDK ==
=====

NVIDIA Release 23.10 (build 72126782)

Copyright (c) 2018-2023, NVIDIA CORPORATION & AFFILIATES. All rights reserved.

Various files include modifications (c) NVIDIA CORPORATION & AFFILIATES. All rights reserved.

This container image and its contents are governed by the NVIDIA Deep Learning Container License.
By pulling and using the container, you accept the terms and conditions of this license:
https://developer.nvidia.com/ngc/nvidia-deep-learning-container-license

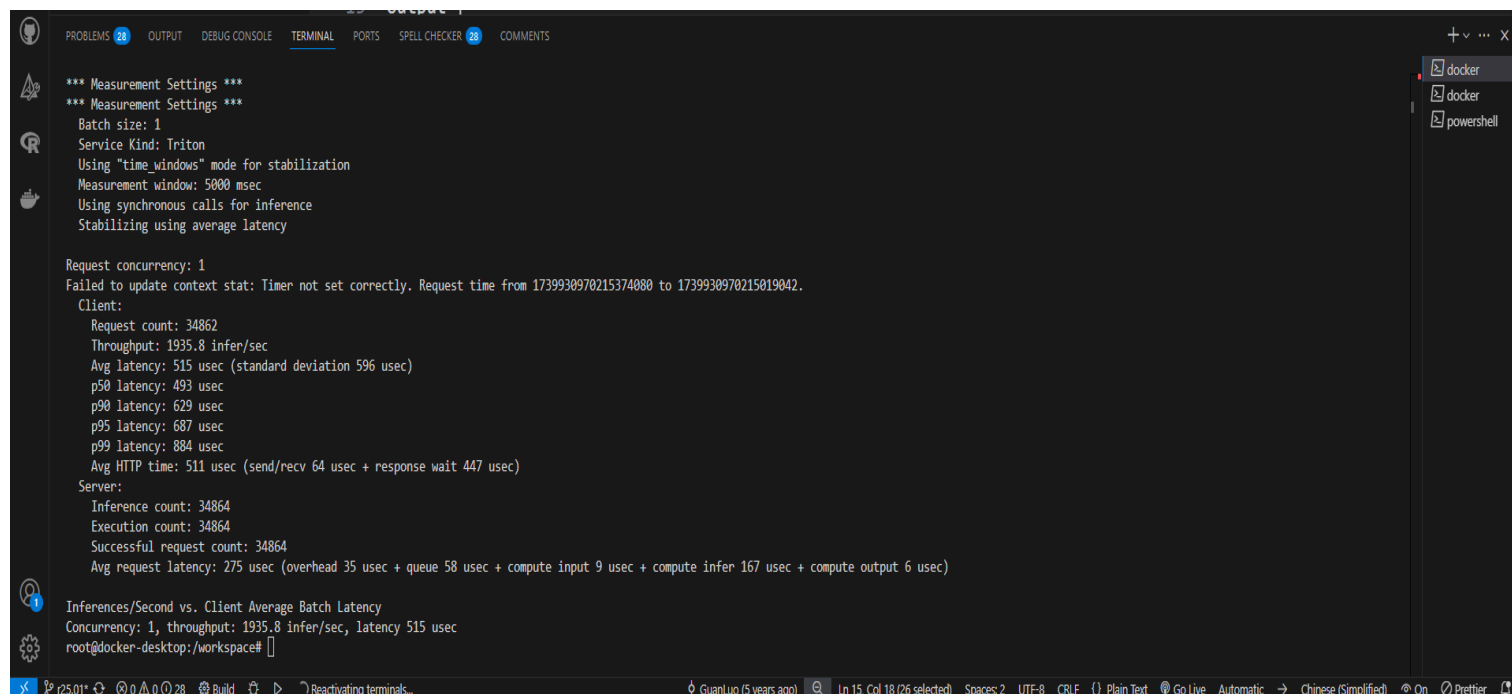
WARNING: The NVIDIA Driver was not detected. GPU functionality will not be available.
Use the NVIDIA Container Toolkit to start this container with GPU support; see
https://docs.nvidia.com/datacenter/cloud-native/ .

Request 0, batch size 1
Image '/workspace/images/mug.jpg':
12.031815 (968) = CUP
9.446401 (504) = COFFEE MUG
9.330780 (899) = WATER JUG

PS D:\SourceCode\ProjectOJT\OJT_TASK3_DEPLOY>
```

3: Performance

3.1 _ Performance_m simple



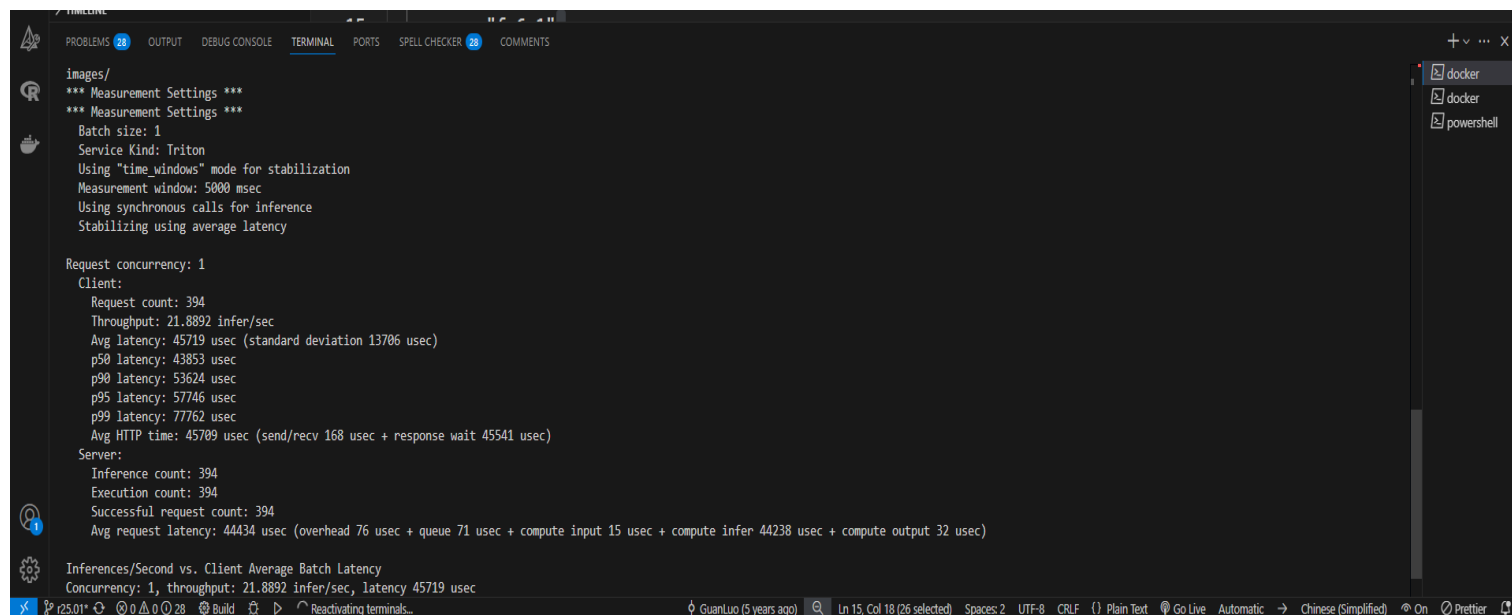
The screenshot shows a VS Code terminal window with the following output:

```
*** Measurement Settings ***
*** Measurement Settings ***
Batch size: 1
Service Kind: Triton
Using "time_windows" mode for stabilization
Measurement window: 5000 msec
Using synchronous calls for inference
Stabilizing using average latency

Request concurrency: 1
Failed to update context stat: Timer not set correctly. Request time from 1739930970215374080 to 1739930970215019042.
Client:
  Request count: 34862
  Throughput: 1935.8 infer/sec
  Avg latency: 515 usec (standard deviation 596 usec)
  p50 latency: 493 usec
  p90 latency: 629 usec
  p95 latency: 687 usec
  p99 latency: 884 usec
  Avg HTTP time: 511 usec (send/recv 64 usec + response wait 447 usec)
Server:
  Inference count: 34864
  Execution count: 34864
  Successful request count: 34864
  Avg request latency: 275 usec (overhead 35 usec + queue 58 usec + compute input 9 usec + compute infer 167 usec + compute output 6 usec)

Inferences/Second vs. Client Average Batch Latency
Concurrency: 1, throughput: 1935.8 infer/sec, latency 515 usec
root@docker-desktop:/workspace#
```

3.2 _ Performance_densenet_onnx



The screenshot shows a VS Code terminal window with the following output:

```
images/
*** Measurement Settings ***
*** Measurement Settings ***
Batch size: 1
Service Kind: Triton
Using "time_windows" mode for stabilization
Measurement window: 5000 msec
Using synchronous calls for inference
Stabilizing using average latency

Request concurrency: 1
Client:
  Request count: 394
  Throughput: 21.8892 infer/sec
  Avg latency: 45719 usec (standard deviation 13706 usec)
  p50 latency: 43853 usec
  p90 latency: 53624 usec
  p95 latency: 57746 usec
  p99 latency: 77762 usec
  Avg HTTP time: 45709 usec (send/recv 168 usec + response wait 45541 usec)
Server:
  Inference count: 394
  Execution count: 394
  Successful request count: 394
  Avg request latency: 44434 usec (overhead 76 usec + queue 71 usec + compute input 15 usec + compute infer 44238 usec + compute output 32 usec)

Inferences/Second vs. Client Average Batch Latency
Concurrency: 1, throughput: 21.8892 infer/sec, latency 45719 usec
```