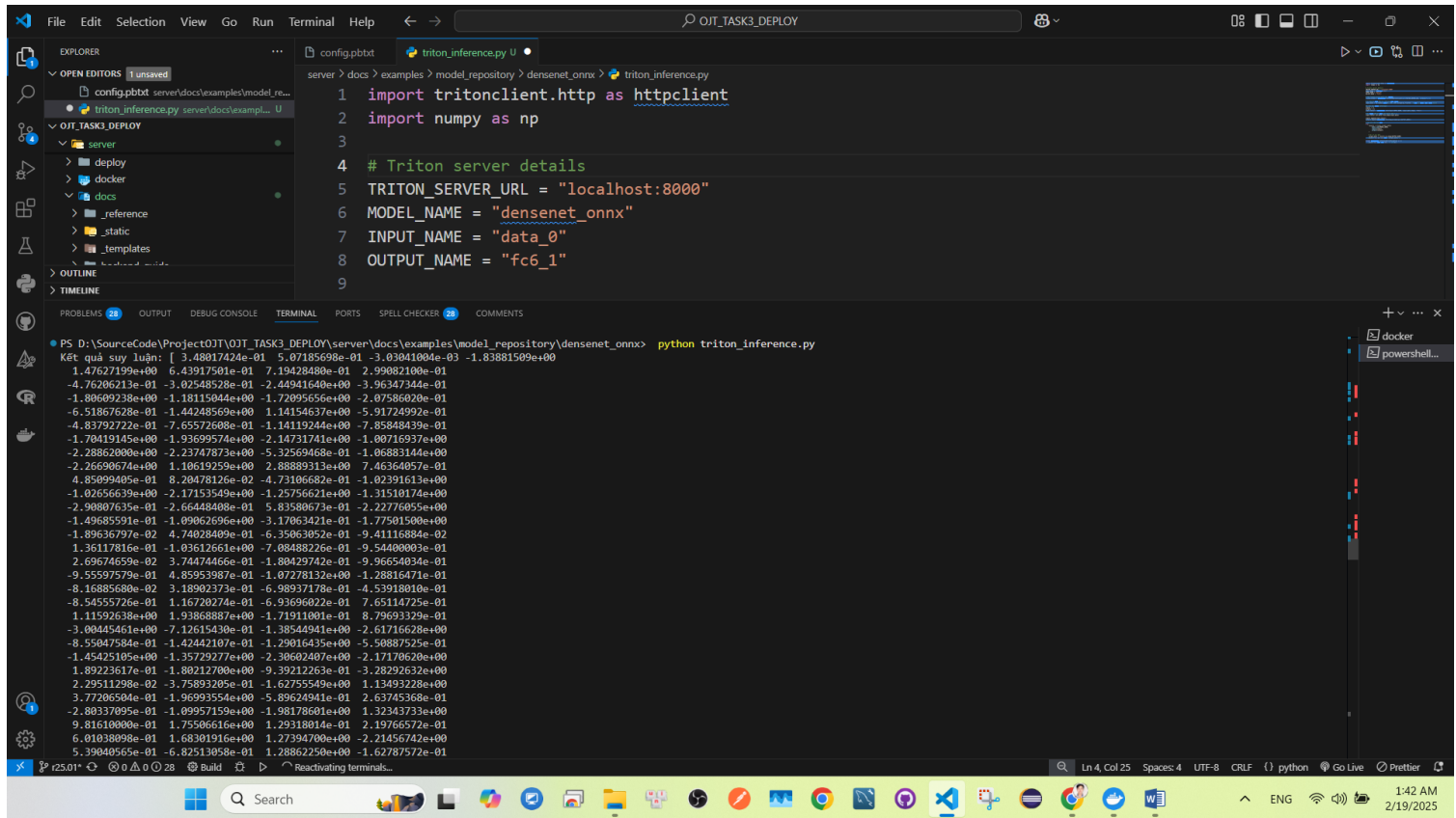


# FINAL - RESULT

## 1. Result Inference



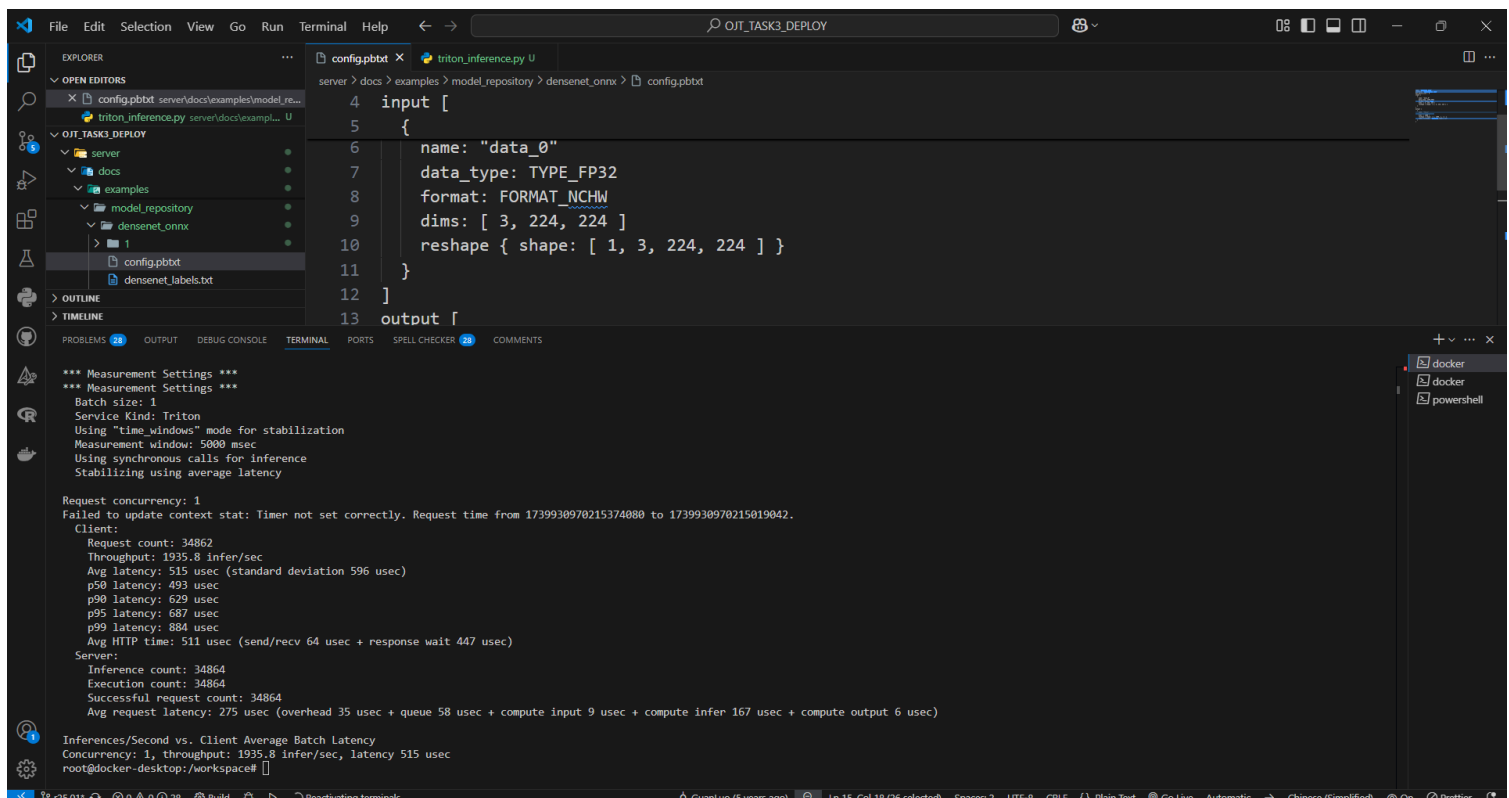
```
config.pbtxt
server > docs > examples > model_repository > densenet_onnx > triton_inference.py

1 import tritonclient.http as httpclient
2 import numpy as np
3
4 # Triton server details
5 TRITON_SERVER_URL = "localhost:8000"
6 MODEL_NAME = "densenet_onnx"
7 INPUT_NAME = "data_0"
8 OUTPUT_NAME = "fc6_1"
9

PS D:\SourceCode\Project\OJT_TASK3_DEPLOY\server\docs\examples\model_repository\densenet_onnx> python triton_inference.py
KEL qua suy luận: [ 3.48017424e-01  5.07185698e-01  3.03041804e-03 -1.83881509e+00
 1.47627159e+00  6.43917501e-01  7.19428480e-01  2.99882100e-01
 4.76206213e-01 -3.02548528e-01 -2.44941640e+00 -3.96347344e-01
 -1.80609238e+00 -1.18115044e+00 -1.7209556e+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  2.88809113e+00  7.46364007e-01
 4.85999405e-01  9.20478126e-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
 -9.5597579e-01  4.85953097e-01 -1.07278132e+00 -1.28816474e-01
 -8.10885680e-02  3.18902373e-01 -6.98937170e-01 -4.53918010e-01
 -8.54555726e-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
 -8.59047584e-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 -5.80624941e-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  1.8862250e+00 -1.62787572e-01
```

## 2: Performance

### 2.1 \_ Performance\_m simple



```
config.pbtxt
server > docs > examples > model_repository > densenet_onnx > config.pbtxt

4 input [
5     {
6         name: "data_0"
7         data_type: TYPE_FP32
8         format: FORMAT_NCHW
9         dims: [ 3, 224, 224 ]
10        reshape { shape: [ 1, 3, 224, 224 ] }
11    }
12 ]
13 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 1739930970215374880 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#
```

## 2.2\_Performance\_densenet\_onnx

