Model and Deployment Analysis Presentation

Yinhao Liu
yl11221
2023-12-14

CONTENTS

Model Summary

Model Summary

Comparison of various models including Previous Model, MobileNetV2, and ResNet101V2.

Analysis includes memory usage, transaction rate, response time, and accuracy.

Aim to identify the model with the best performance based on business requirements.

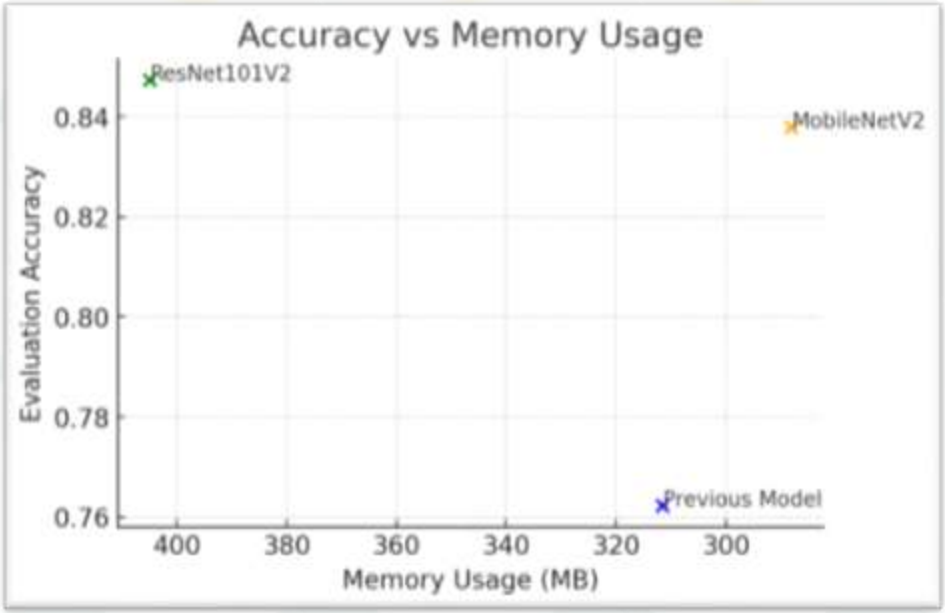
Metric	Previous Model	MobileNetV2	ResNet101V2
Transactions	465	461	97
Availability (%)	100.00	100.00	8.58
Elapsed Time (s)	29.49	29.54	20.29
Response Time (s)	0.63	0.63	1.67
Transaction Rate (tps)	15.77	15.61	4.78
Concurrency	9.92	9.86	7.98
Successful Transactions	465	461	97
Failed Transactions	0	0	1033
Longest Transaction (s)	1.82	2.00	2.68
Shortest Transaction (s)	0.25	0.39	0.54
Disk space usage (MB)	9.27	12.4 MB	172 MB
Memory usage	311.5MB	288.1MB	404.9MB
Evaluation Accuracy	0.7622	0.8378	0.8473

Accuracy vs Memory Usage

Accuracy vs Memory Usage

Visualizes the trade-off between accuracy and memory usage.

Highlights MobileNetV2's efficiency in terms of high accuracy with low memory usage.

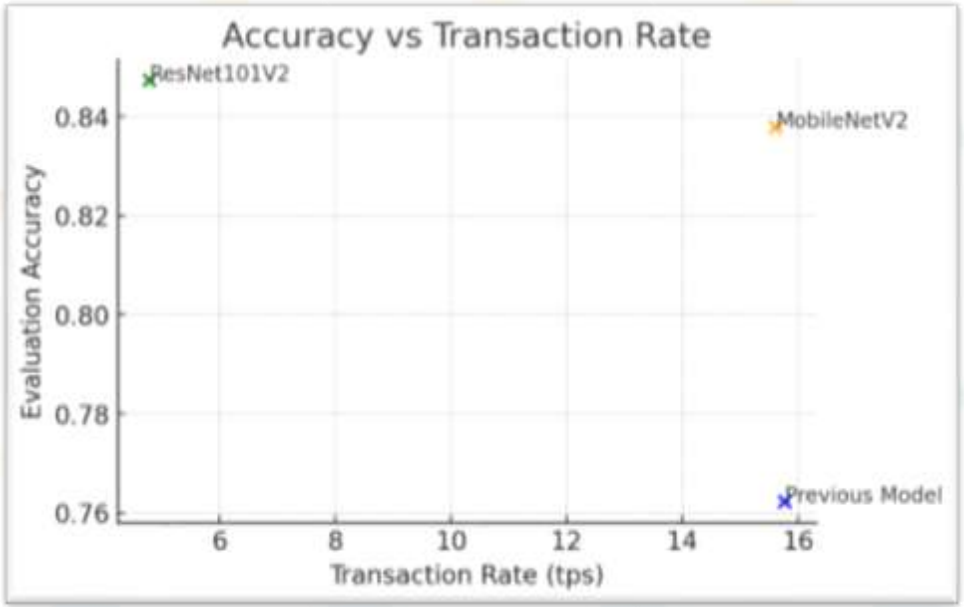


Accuracy vs Transaction Rate

Accuracy vs Transaction Rate

Shows each model's ability to handle transactions at different rates.

ResNet101V2 (Dynamic Scaling) excels in high-load scenarios.

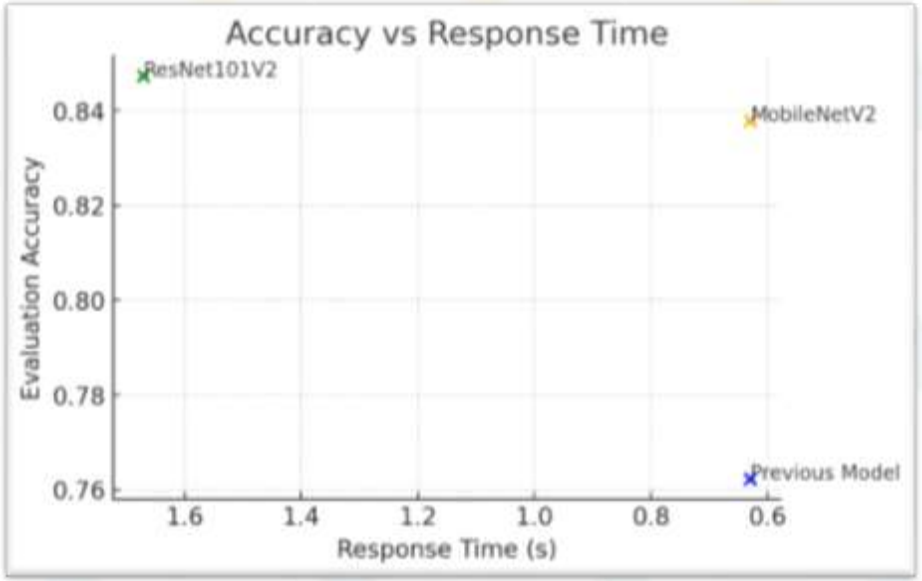


Accuracy vs Response Time

Accuracy vs Response Time

Compares models based on their response time and accuracy.

Previous Model offers quick responses, ideal for real-time applications.



Deployment Options Summary

Deployment Options Summary

Model / Metric	Transactions	Availability (%)	Elapsed Time (s)	Data Transferred	Response Time (s)	Transaction Rate (tps)	Throughput	Concurrency	Successful Transactions	Failed Transactions	Longest Transaction (s)	Shortest Transaction (s)
Pretrained Model	465.0	100.0	29.49	0.01	0.63	15.77	0.0	9.92	465.0	0.0	1.82	0.25
MobileNetV2	461.0	100.0	29.54	0.01	0.63	15.61	0.0	9.86	461.0	0.0	2.0	0.39
ResNet101V2 (Dynamic Scaling)	8346.0	100.0	359.8	0.25	0.43	23.2	0.0	9.99	8346.0	0.0	1.49	0.16
ResNet101V2 (Load Balancing)	542.0	100.0	29.66	0.02	0.54	18.27	0.0	9.91	542.0	0.0	1.46	0.16
ResNet101V2 (Default)	97.0	8.58	20.29	0.0	1.67	4.78	0.0	7.98	97.0	1033.0	2.68	0.54

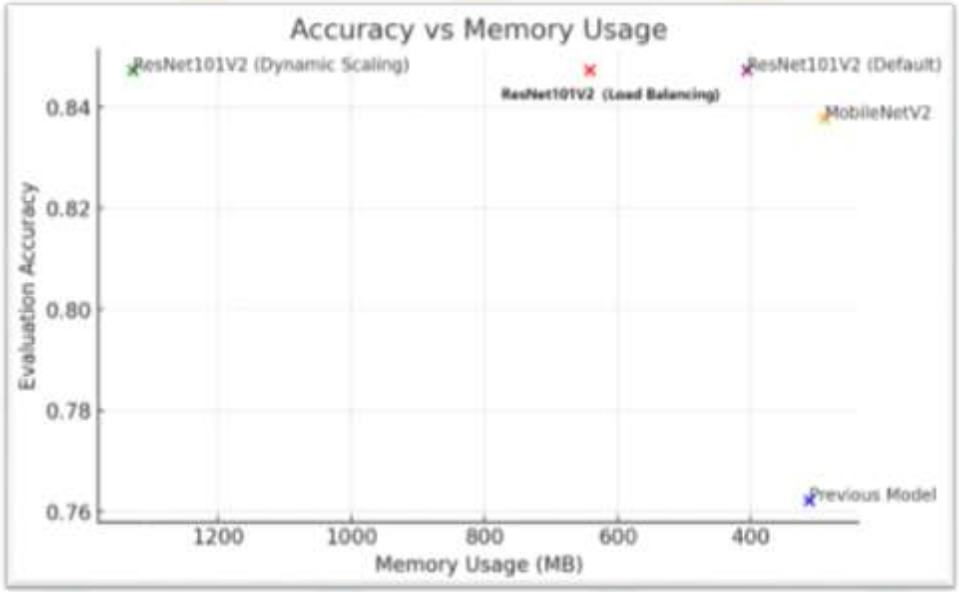
Overview of different deployment strategies for ResNet101V2.

Includes dynamic scaling, load balancing, and default deployment.

Accuracy vs Memory Usage II

Accuracy vs Memory Usage II

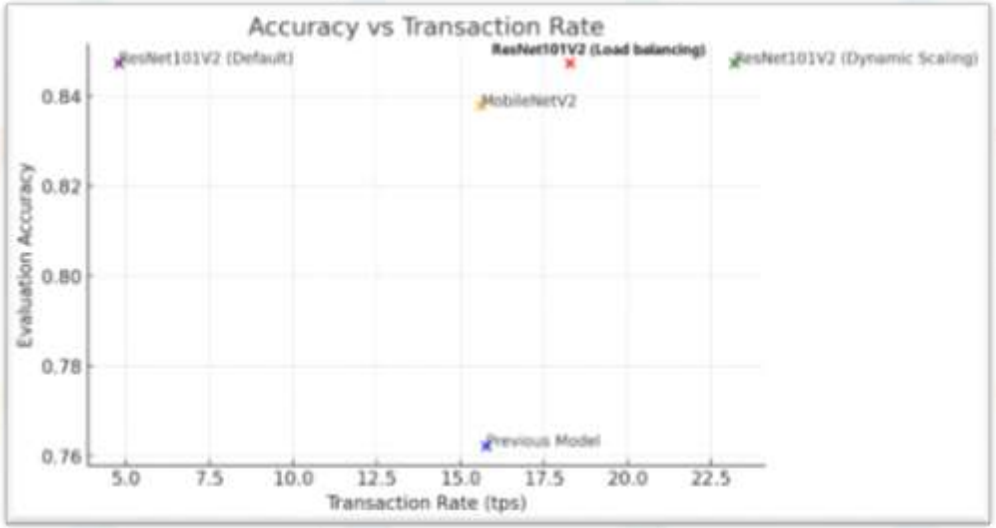
- MobileNetV2 provides high accuracy with low memory usage.
- ResNet101V2 (Default) uses the most memory without a corresponding increase in accuracy.
- The Previous Model is efficient in memory usage but has lower accuracy.
- ResNet101V2 (Load Balancing) offers a good balance but is outperformed by MobileNetV2 in terms of efficiency.



Accuracy vs Transaction Rate II

Accuracy vs Transaction Rate II

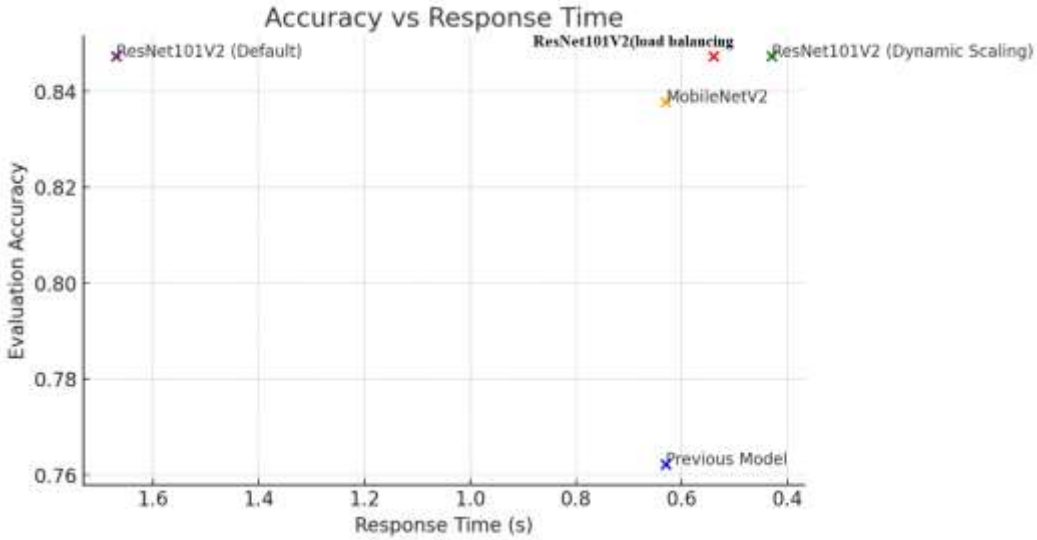
- ResNet101V2 (Dynamic Scaling) achieves the highest transaction rate and maintains high accuracy.
- MobileNetV2 shows competitive accuracy with a good transaction rate, striking a balance between performance and efficiency.
- The Previous Model, while still accurate, manages fewer transactions per second, potentially limiting its suitability for high-traffic scenarios.
- ResNet101V2 (Default) has lower transaction rates, suggesting possible performance constraints under default settings.



Accuracy vs Response Time II

Accuracy vs Response Time II

- ResNet101V2 (Dynamic Scaling) exhibits the best performance with the highest accuracy and the fastest response time.
- MobileNetV2 also shows strong performance, with high accuracy and a quick response time, suggesting it is well-optimized for real-time applications.
- The Previous Model has a slightly slower response time, which might affect performance in time-sensitive tasks.
- ResNet101V2 (Default) has the slowest response time, which could be a significant drawback for applications where latency is critical, despite its high accuracy.



Conclusion and Recommendations

Conclusion and Recommendations

- If the priority is to minimize costs and conserve resources, MobileNetV2 might be the best choice.
- For services that require high throughput and can accommodate higher costs, ResNet101V2 (Dynamic Scaling) would be appropriate.
- When there is a need for a balance between performance and cost, ResNet101V2 (Load Balancing) might be the optimal path.
- If the application can handle variability in performance and potential stability issues for the sake of high accuracy, then ResNet101V2 (Default) could be considered.

THE END

THANKS