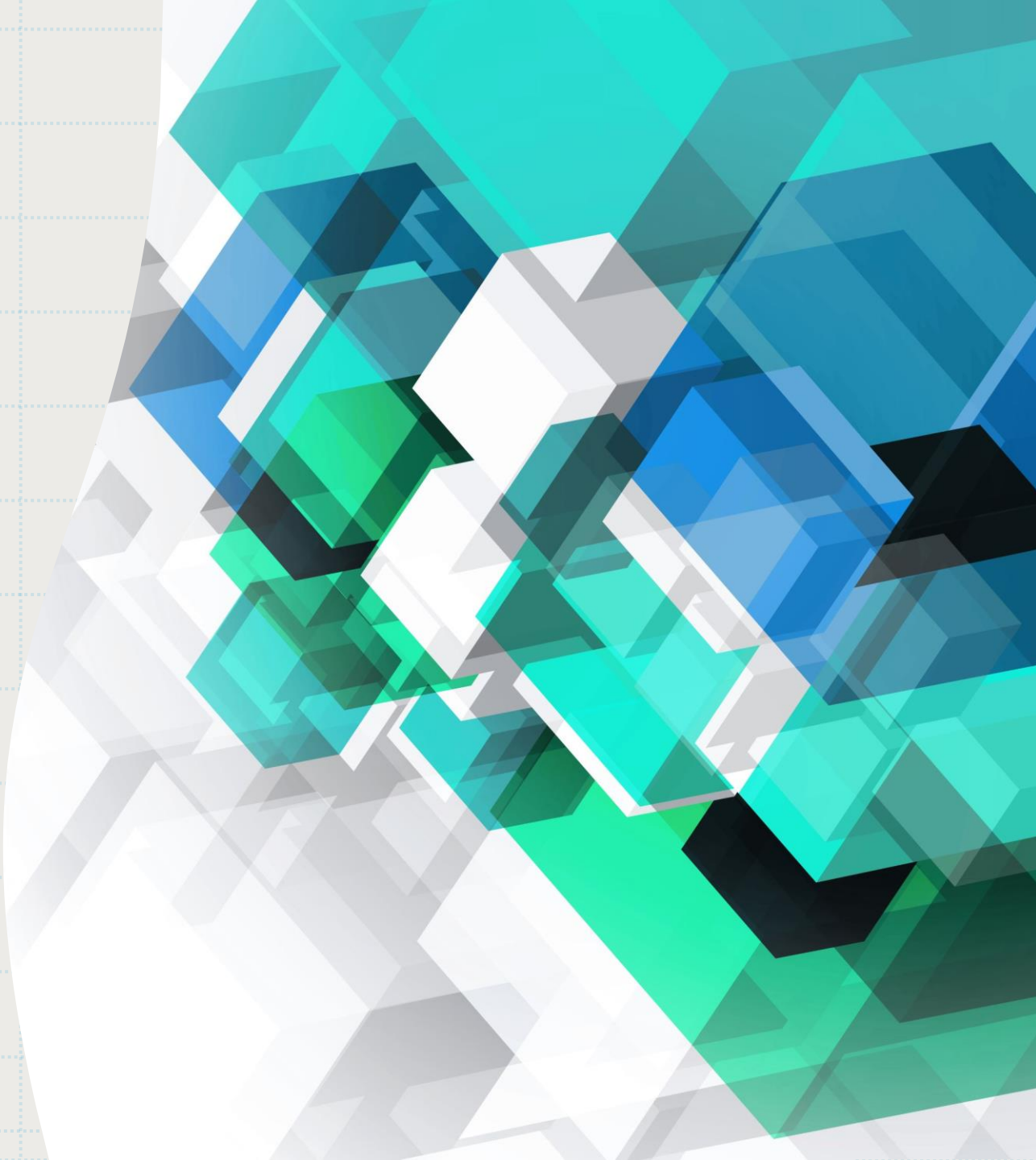
From Training to Deployment: Improve the Food Classification

Yihua Yang

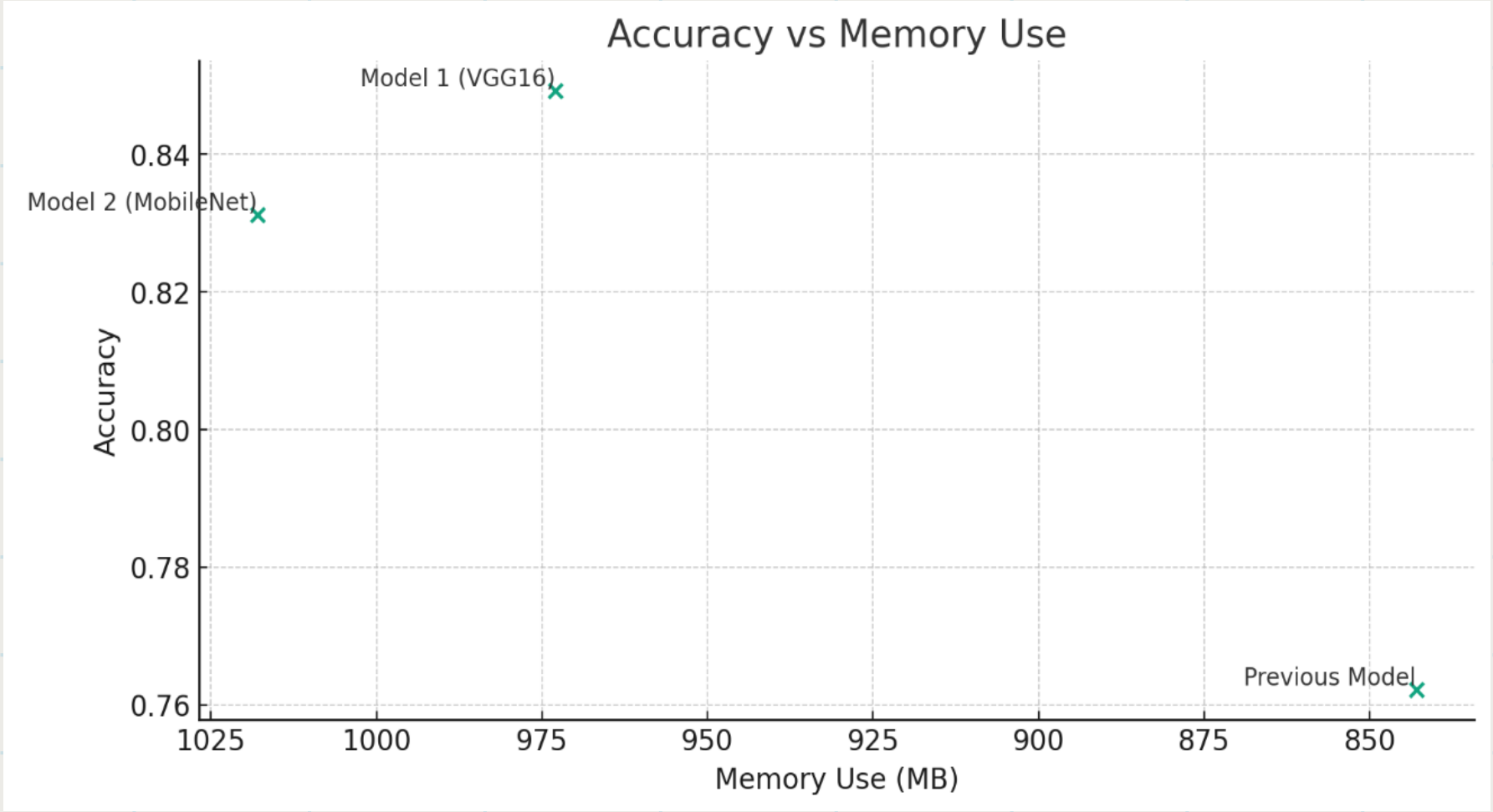
yy5028@nyu.edu



Model Summary

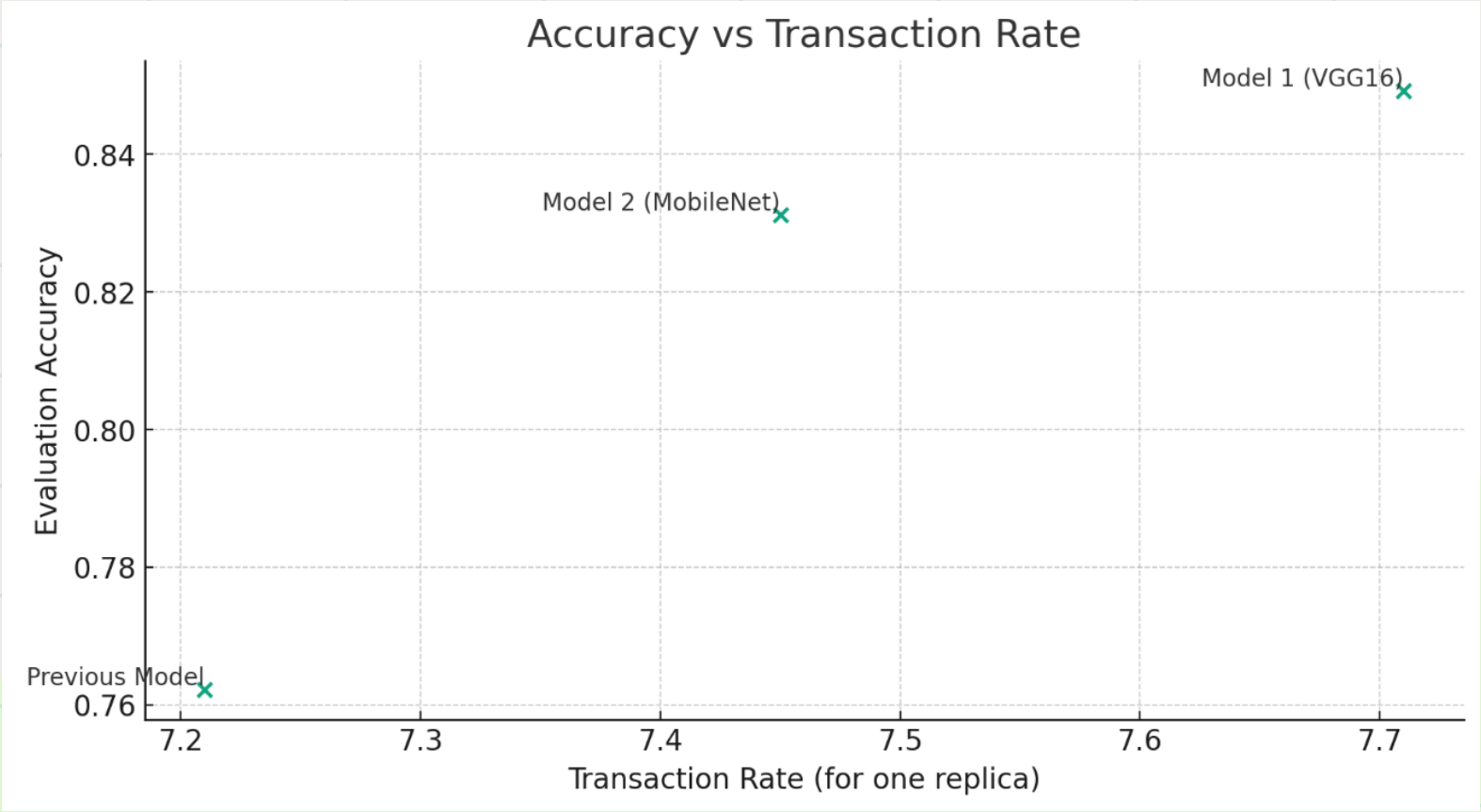
	Previous Model	*Model 1 (VGG16)*	*Model 2 (MobileNet)*
Overall Accuracy	0.7622	0.8492	0.8312
Disk Space	9.27 MB	110 MB	63.1 MB
Response Time	1.17	1.08	1.11
Bad Classification	in 6 classes	in 3 classes	in 5 classes
Memory Use	843 MB	973 MB	1018 MB

Accuracy vs. Memory Use (single replica)

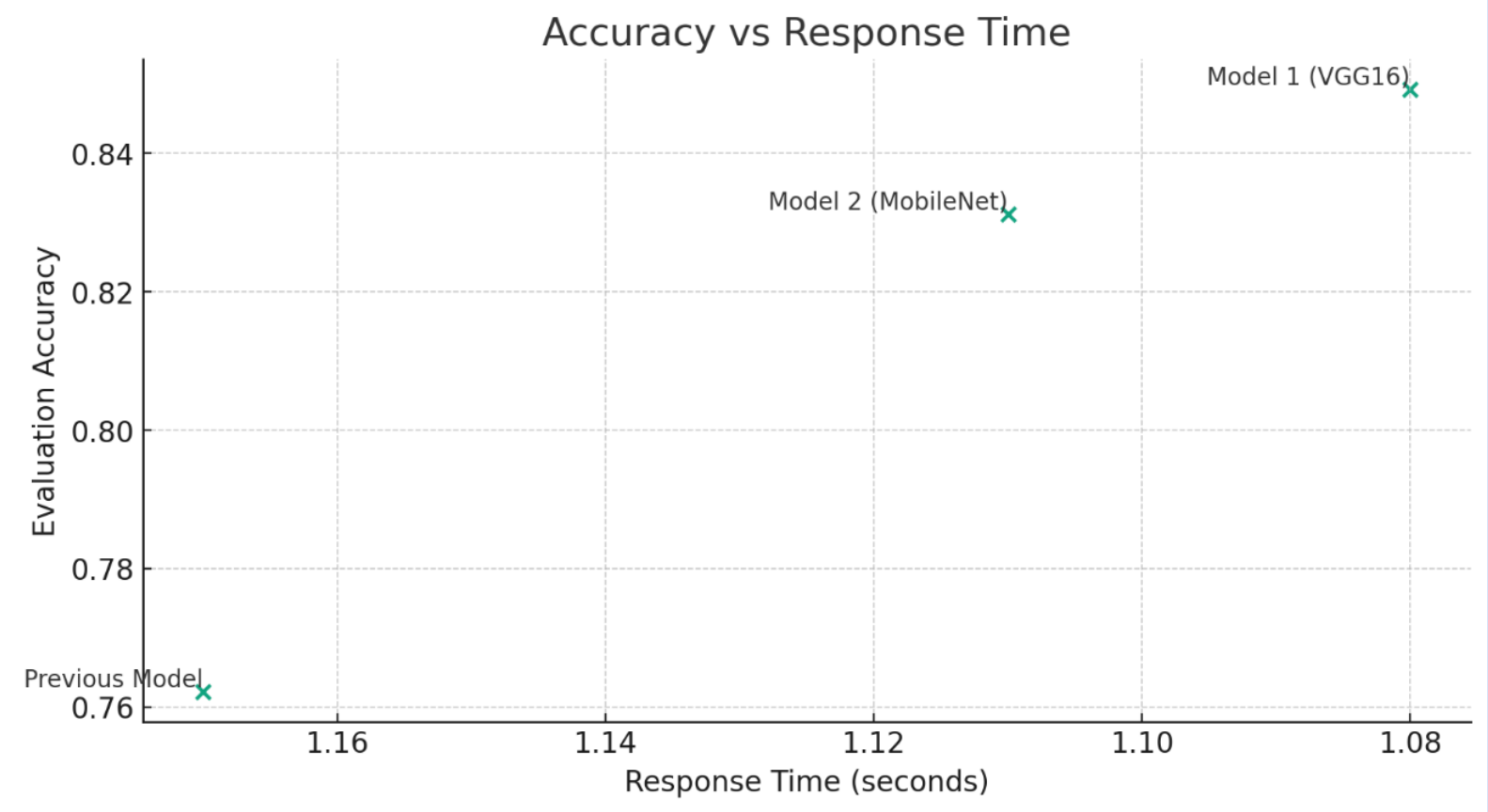


- Model 1 is the one with best accuracy but consumes more memory
- Model 2 has lower memory use and lower accuracy

Accuracy vs. Transaction Rate (single replica)



Accuracy vs. Response Time (single replica)



Model 1 has the best performance in Accuracy vs. Response Time

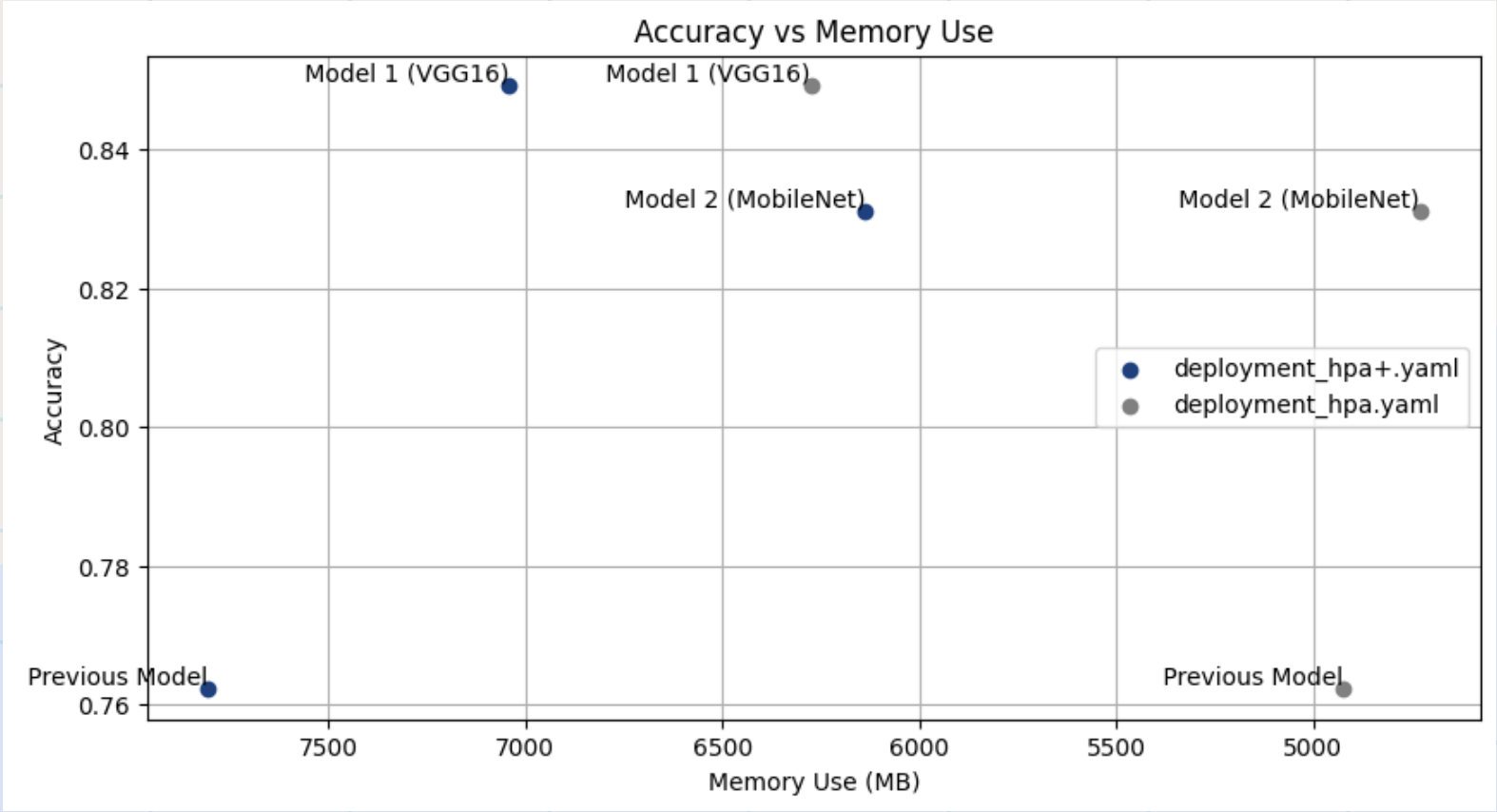
Deployment_hpa vs. Deployment_hpa+

Using deployment_hpa.yaml				
	CPU Usage (Cores)	Memory Usage (KB)	Response Time (Sec)	Transaction Rate
Previous Model	2.379	5041017	0.37	13.28
Model 1	2.425	6425222	0.46	13.82
Model 2	2.361	4841846	0.48	12.87

Using deployment_hpa+.yaml				
	CPU Usage (Cores)	Memory Usage (KB)	Response Time (Sec)	Transaction Rate
Previous Model	2.504	7994066	0.42	14.14
Model 1	2.499	7210982	0.43	14.15
Model 2	2.512	6284285	0.42	14.22

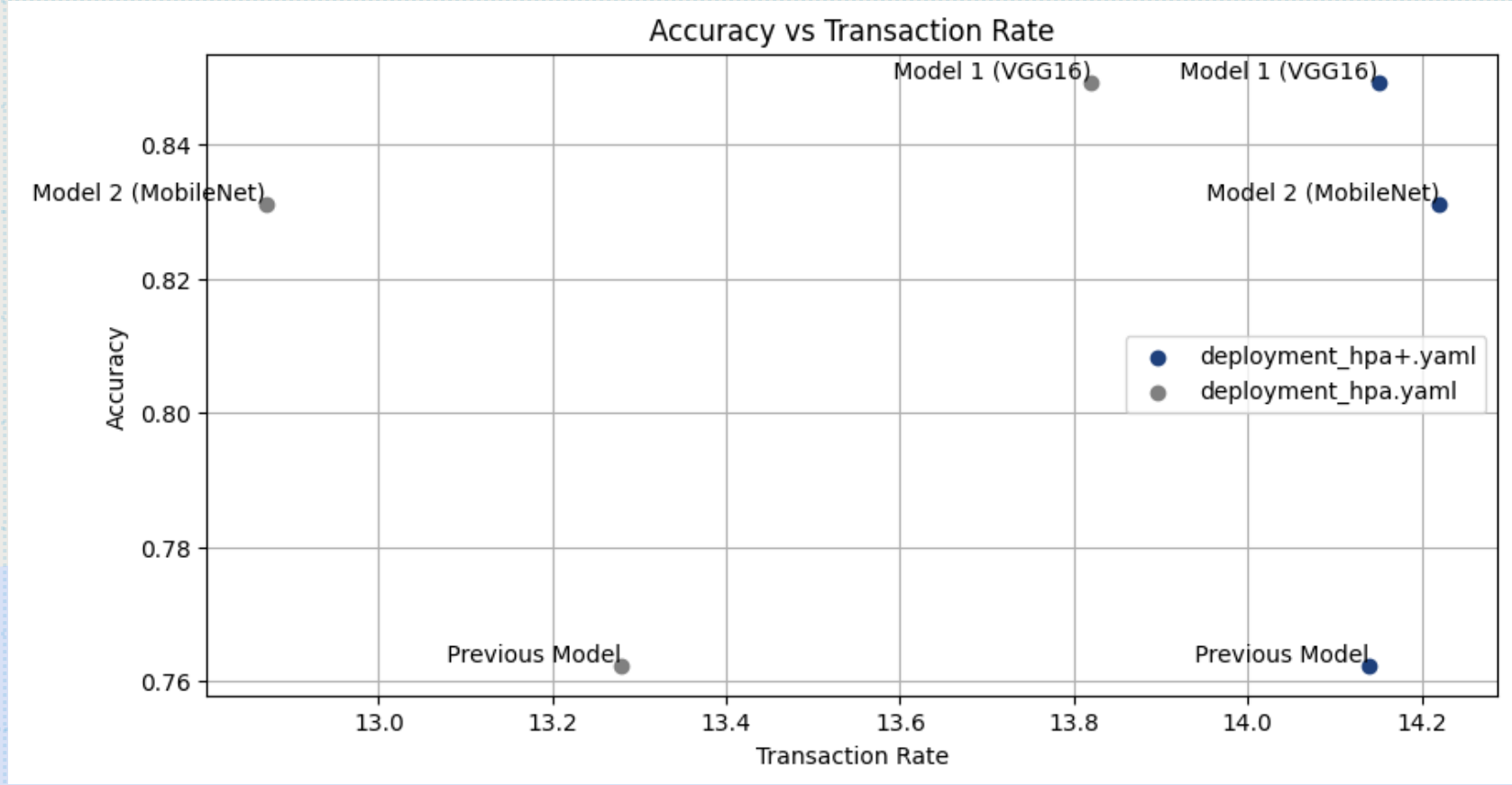
Compare _hpa and _hpa+: Accuracy vs. Memory Use

- Best Memory Use: model 2
- Best Accuracy: model 1



Compare _hpa and _hpa+: Accuracy vs. Transaction Rate

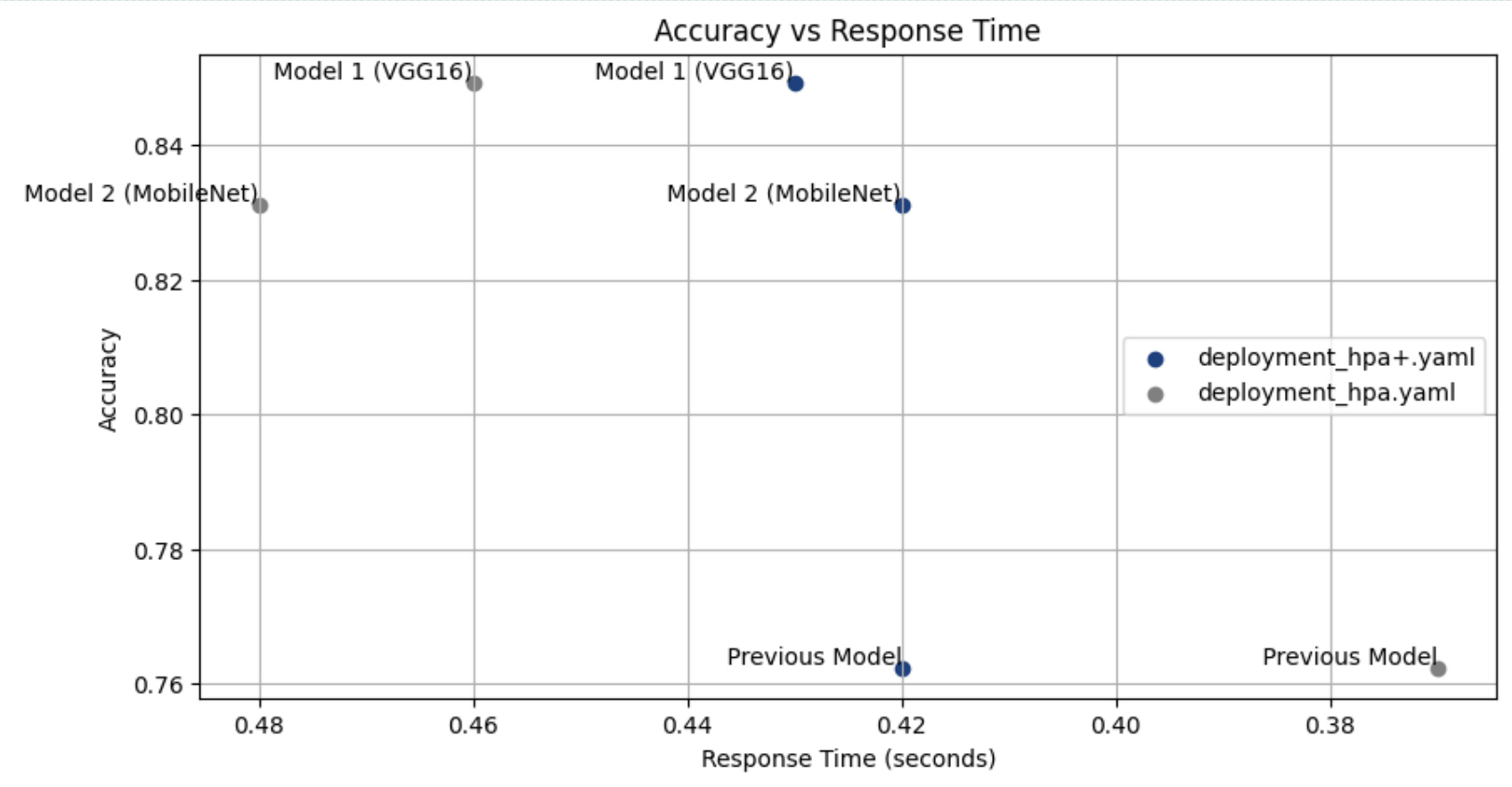
- Best Accuracy: model 1
- Best Transaction Rate: model 2



Compare _hpa and _hpa+: Accuracy vs. Response Time

Best Response Time: Previous Model

Best Accuracy: VGG16



Summary

- **Improvement in Accuracy:**
- From 0.7622 to 0.8312 (based on MobileNetV2) / 0.8492 (based on VGG16)
- **Improvement in Transaction Rate:**
- From 13.28 to 14.15 (VGG16&hpa+.yaml) /14.22 (MobileNetV2&hpa+.yaml)
- **Improvement in Response Time:**
- From 0.48 (MobileNetV2&hpa) to 0.42 (MobileNetV2&hpa+)

Choose which one?

Accuracy -> Model 1 (VGG16)

Memory Use -> Model 2 (MobileNetV2)