ECE 361E: Machine Learning and Data Analytics for Edge AI

Homework 3 Solution

Team Members and Contributions:

Ryane Li: Worked on Problem 1, including creating the VGG16 model architecture, training VGG11, VGG16, and MobileNet-v1 models on Lonestar6, generating the required metrics for Table 1, and creating the test accuracy comparison plot.

Resul Ovezov: Worked on Problems 2 and 3, including converting models to ONNX format, deploying VGG11, VGG16, and MobileNet-v1 on RaspberryPi 3B+ and Odroid MC1 edge devices, collecting inference metrics, measuring power consumption and temperature variations.

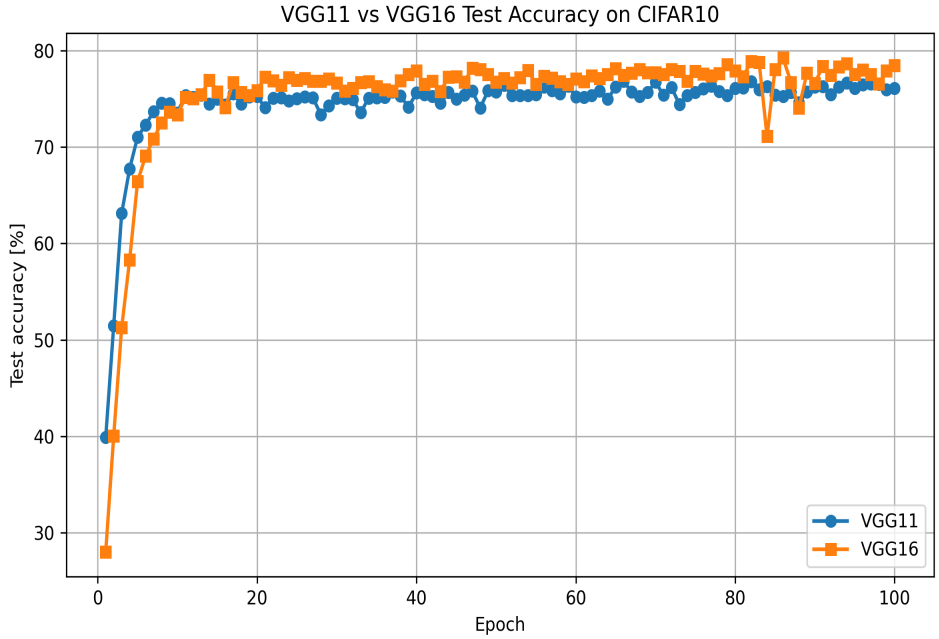
Problem 1: PyTorch Evaluation of VGG Models

Question 2: Training Metrics

Table 1: Training Metrics for VGG11, VGG16, and MobileNet-v1

Model	Train Acc %	Test Acc %	Train Time (s)	Params (M)	FLOPs (M)	GPU Mem (MB)
VGG11	99.07	76.08	1468.6	9.75	306.6	939
VGG16	98.35	78.46	1622.6	15.25	627.5	2037
MOBILENET-V1	99.37	78.47	1754.2	3.22	96.0	1263

Question 3: Test Accuracy Comparison



Analysis and Comparison (VGG11 vs VGG16):

• **Accuracy vs. epochs:** VGG11 reaches higher accuracy in the very early epochs, but VGG16 quickly catches up and ultimately achieves better final test accuracy. After 100 epochs, VGG11 ends at 76.08% test accuracy while VGG16 reaches 78.46%, so VGG16 provides about +2.4 percentage points better generalization.

• **Training accuracy and overfitting:** Both models achieve very high training accuracy (VGG11: 99.07%, VGG16: 98.35%). The gap between train and test accuracy indicates some overfitting for both, but the gap is not dramatically worse for VGG16, even though it is deeper.

• **Training time:** VGG16 takes longer to train (1622.58 s) than VGG11 (1468.63 s), roughly 10% more wall-clock time for the same number of epochs on the same hardware.

• **Model size and FLOPs:** VGG16 is substantially heavier: VGG11 has 9.75M trainable parameters while VGG16 has 15.25M (about 1.6× more). VGG11 requires 306.6M FLOPs per forward pass, while VGG16 requires 627.5M FLOPs (about 2× more compute).

• **GPU memory usage:** VGG16 also uses more GPU memory during training (2037 MB) than VGG11 (939 MB), which matters if GPU memory is a bottleneck.

Conclusion: VGG16 provides slightly better test accuracy (about 2-3 percentage points) but at the cost of roughly 2× FLOPs, 1.6× parameters, higher GPU memory usage, and slightly longer training time. If maximum accuracy is the only goal and compute/memory are plentiful, VGG16 is preferable. However, in edge- or resource-constrained settings—where training and inference cost matter as much as accuracy—VGG11 is more attractive because it is significantly cheaper while achieving only a modestly

lower test accuracy. In this homework context, where we care about efficiency on edge devices, we would generally prefer VGG11.

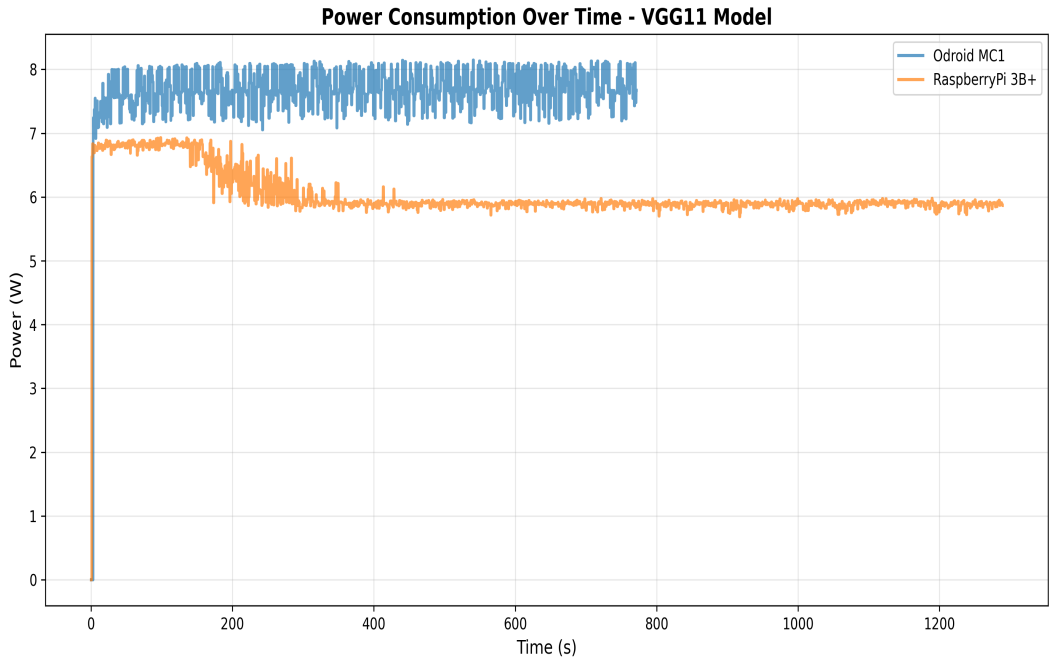
Problem 2: Deployment on Edge Devices Using ONNX

Question 2: Inference Metrics on Edge Devices

Table 2: Inference Performance on Edge Devices

	Total Inference Time (s)		RAM Memory (MB)		Accuracy (%)	
Model	MC1	RaspberryPi	MC1	RaspberryPi	MC1	RaspberryPi
VGG11	771.48	1289.28	331.00	166.00	76.10	76.10
VGG16	1115.73	1801.56	355.00	185.00	78.47	78.47
MOBILENET	542.76	1018.98	301.00	131.00	78.49	78.49

Question 3: Power Consumption and Temperature Analysis



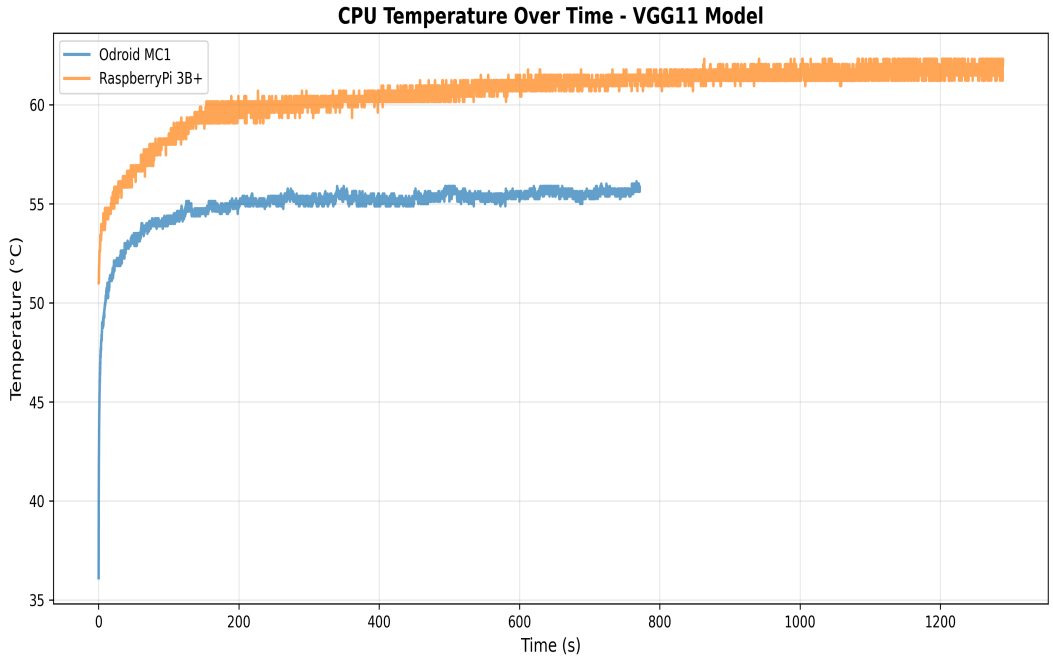


Table 3: Energy Consumption on Edge Devices

Model	MC1 Total Energy (J)	RaspberryPi Total Energy (J)
VGG11	5911.8795	7799.9015
VGG16	9698.1863	10698.0575
MOBILENET	3600.3522	6022.7317

Performance Comparison and Analysis:

Based on the VGG11 and VGG16 models for the RaspberryPi and the MC1, the MC1 is best for inference. This is because while RaspberryPi takes up half the memory space, its inference time and total energy used is a great amount larger. The total inference time for MC1 was 771.48s while the RaspberryPi took 1289.28s for the VGG11. The large difference between the inference times also resulted in the RaspberryPi using more energy overall even though the power consumption rate for the MC1 is higher than the Pi.

Problem 3: MobileNet-v1 on Edge Devices

Questions 1-2: MobileNet-v1 was trained on CIFAR10 using Lonestar6 and deployed on both edge devices. The results have been incorporated into the extended versions of Tables 1, 2, and 3 shown above.

BONUS Question 3: Comprehensive Model Analysis

MobileNet drastically outperforms all the other models on both the MC1 and RaspberryPi.

For the MC1:

MobileNet latency is the fastest at 54.28 ms per image (vs 77.15 ms for VGG11, 111.57 ms for VGG16). It is the most energy efficient at 0.3600 J per image (vs 0.5912 J for VGG11, 0.9698 J for VGG16). Accuracy is the highest at 78.49% and has the smallest amount of RAM usage with only 301 MB.

For the RaspberryPi 3B+:

MobileNet latency is the fastest at 101.90 ms per image (vs 128.93 ms for VGG11, 180.16 ms for VGG16). It is the most energy efficient at 0.6023 J per image (vs 0.7800 J for VGG11, 1.0698 J for VGG16). Accuracy is the highest at 78.49% and has the smallest amount of RAM usage with only 131 MB.