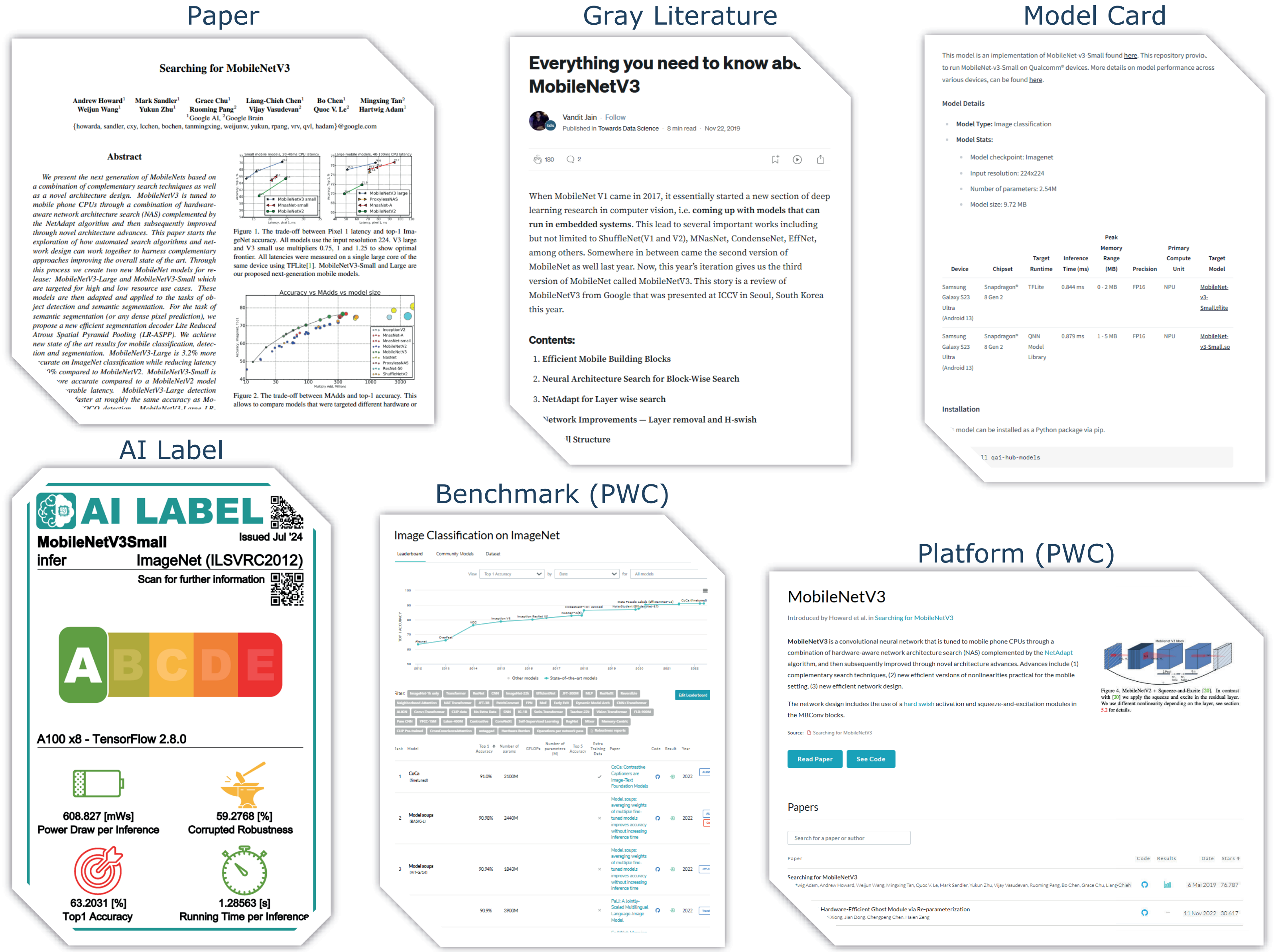

Paper
Searching for MobileNetV3
Andrew Howard[1] Mark Sandler[1] Grace Chu[1] Liang-Chieh Chen[1] Bo Chen[1] Mingxing Tan[2]
Weijun Wang[1] Yukun Zhu[1] Ruoming Pang[2] Vijay Vasudevan[2] Quoc V. Le[2] Hartwig Adam[1]
[1]Google AI, [2]Google Brain
{howarda, sandler, cxy, lcchen, bochen, tanmingxing, weijunw, yukun, rpang, vrv, qvl, hadam}@google.com
Abstract
We present the next generation of MobileNets based on a combination of complementary search techniques as well as a novel architecture design. MobileNetV3 is tuned to mobile phone CPUs through a combination of hardware-aware network architecture search (NAS) complemented by the NetAdapt algorithm and then subsequently improved through novel architecture advances. This paper starts the exploration of how automated search algorithms and network design can work together to harness complementary approaches improving the overall state of the art. Through this process we create two new MobileNet models for release: MobileNetV3-Large and MobileNetV3-Small which are targeted for high and low resource use cases. These models are then adapted and applied to the tasks of object detection and semantic segmentation. For the task of semantic segmentation (or any dense pixel prediction), we propose a new efficient segmentation decoder Lite Reduced Atrous Spatial Pyramid Pooling (LR-ASPP). We achieve new state of the art results for mobile classification, detection and segmentation. MobileNetV3-Large is 3.2% more accurate on ImageNet classification while reducing latency
Figure 1. The trade-off between Pixel 1 latency and top-1 ImageNet accuracy. All models use the input resolution 224. V3 large and V3 small use multipliers 0.75, 1 and 1.25 to show optimal frontier. All latencies were measured on a single large core of the same device using TFLite[1]. MobileNetV3-Small and Large are our proposed next-generation mobile models.
Accuracy vs MAdds vs model size
Figure 2. The trade-off between MAdds and top-1 accuracy. This allows to compare models that were targeted different hardware or
Gray Literature
Everything you need to know abo
MobileNetV3
Vandit Jain · Follow
Published in Towards Data Science · 8 min read · Nov 22, 2019
When MobileNet V1 came in 2017, it essentially started a new section of deep learning research in computer vision, i.e. coming up with models that can run in embedded systems. This lead to several important works including but not limited to ShuffleNet(V1 and V2), MNasNet, CondenseNet, EffNet, among others. Somewhere in between came the second version of MobileNet as well last year. Now, this year's iteration gives us the third version of MobileNet called MobileNetV3. This story is a review of MobileNetV3 from Google that was presented at ICCV in Seoul, South Korea this year.
Contents:
1. Efficient Mobile Building Blocks
2. Neural Architecture Search for Block-Wise Search
3. NetAdapt for Layer wise search
Network Improvements — Layer removal and H-swish
ll Structure
Model Card
This model is an implementation of MobileNet-v3-Small found here. This repository provides to run MobileNet-v3-Small on Qualcomm® devices. More details on model performance across various devices, can be found here.
Model Details
Model Type: Image classification
Model Stats:
Model checkpoint: Imagenet
Input resolution: 224x224
Number of parameters: 2.54M
Model size: 9.72 MB
Device | Chipset | Target Runtime | Inference Time (ms) | Peak Memory Range (MB) | Precision | Primary Compute Unit | Target Model
Samsung Galaxy S23 Ultra (Android 13) | Snapdragon® 8 Gen 2 | TFLite | 0.844 ms | 0 - 2 MB | FP16 | NPU | MobileNet-v3-Small.tflite
Samsung Galaxy S23 Ultra (Android 13) | Snapdragon® 8 Gen 2 | QNN Model Library | 0.879 ms | 1 - 5 MB | FP16 | NPU | MobileNet-v3-Small.so
Installation
model can be installed as a Python package via pip.
ll qai-hub-models
AI Label
AI LABEL
MobileNetV3Small
Issued Jul '24
infer
ImageNet (ILSVRC2012)
Scan for further information
A B C D E
A100 x8 - TensorFlow 2.8.0
608.827 [mWs]
Power Draw per Inference
59.2768 [%]
Corrupted Robustness
63.2031 [%]
Top1 Accuracy
1.28563 [s]
Running Time per Inference
Benchmark (PWC)
Image Classification on ImageNet
Leaderboard Community Models Dataset
Rank | Model | Top 1 Accuracy | Number of params | Year
1 | CoCa (finetuned) | 91.0% | 2100M | 2022
2 | Model soups (BASIC-L) | 90.98% | 2440M | 2022
3 | Model soups (ViT-G/14) | 90.94% | 1843M | 2022
| | 90.9% | 3900M | 2022
Platform (PWC)
MobileNetV3
Introduced by Howard et al. in Searching for MobileNetV3
MobileNetV3 is a convolutional neural network that is tuned to mobile phone CPUs through a combination of hardware-aware network architecture search (NAS) complemented by the NetAdapt algorithm, and then subsequently improved through novel architecture advances. Advances include (1) complementary search techniques, (2) new efficient versions of nonlinearities practical for the mobile setting, (3) new efficient network design.
The network design includes the use of a hard swish activation and squeeze-and-excitation modules in the MBConv blocks.
Source: Searching for MobileNetV3
Read Paper
See Code
Papers
Search for a paper or author
Searching for MobileNetV3
6 Mai 2019 76.787
Hardware-Efficient Ghost Module via Re-parameterization
11 Nov 2022 30.617