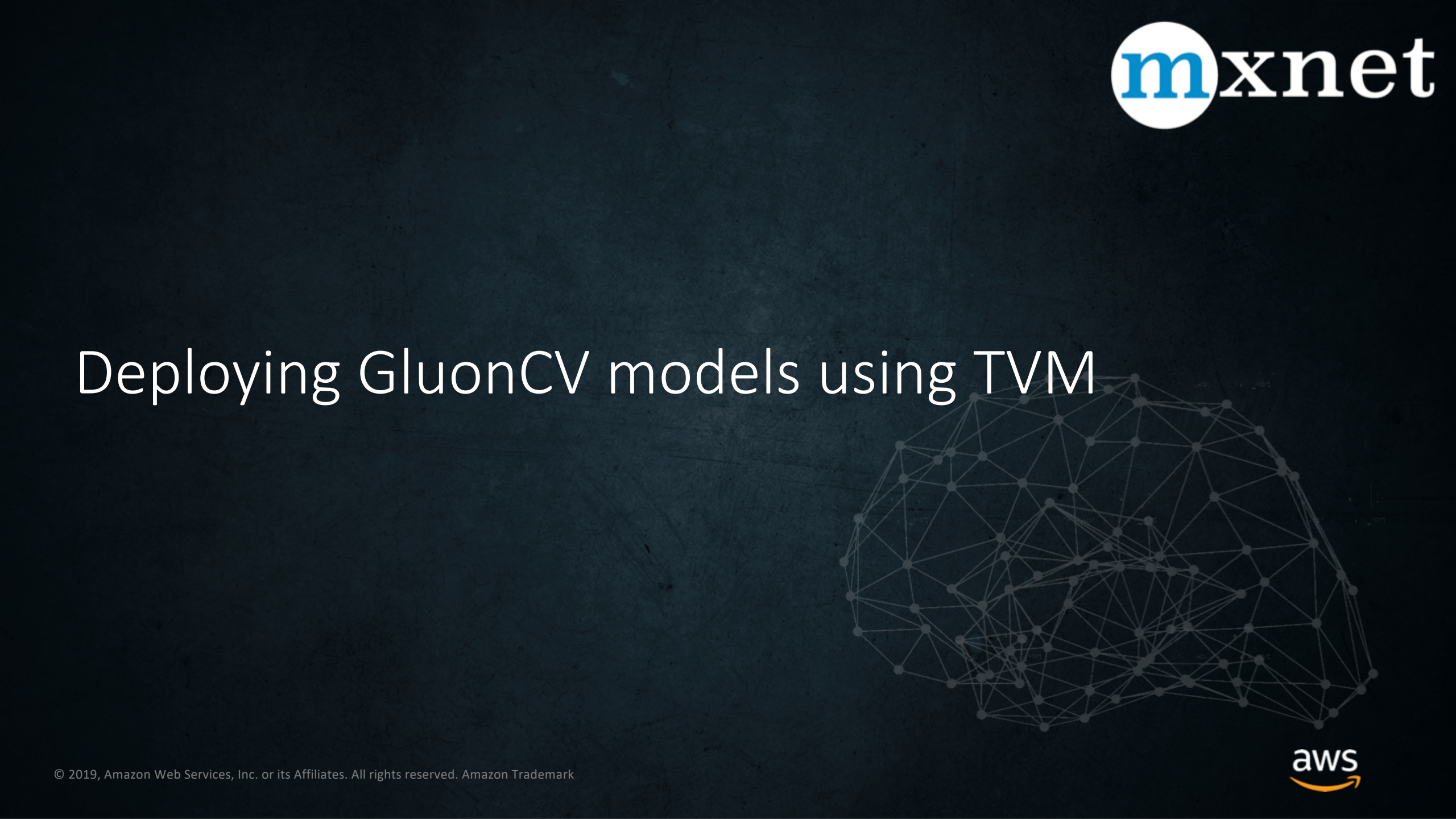
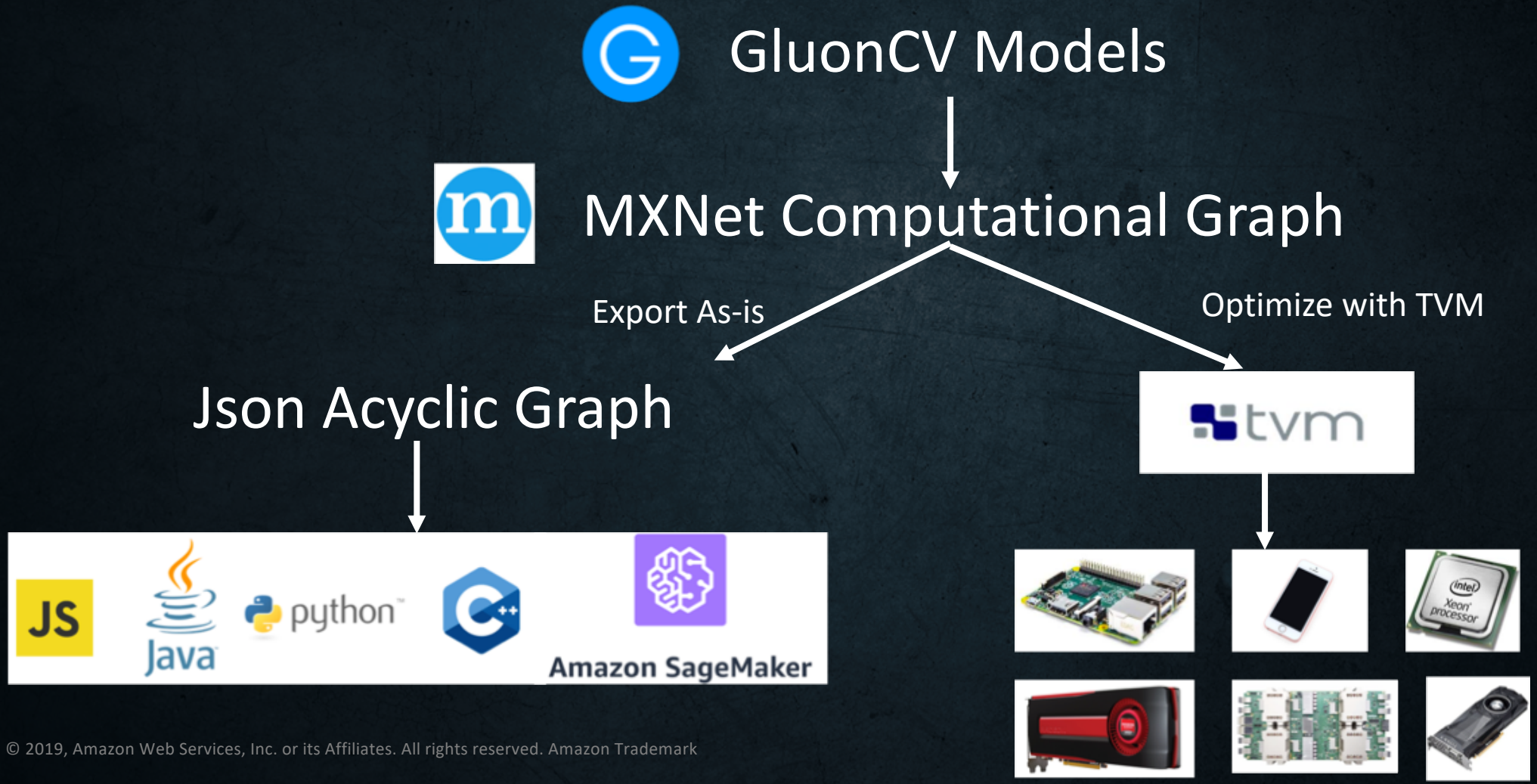


Deploying GluonCV models using TVM

Deploy GluonCV Models



Deploy GluonCV Models

A Unified Optimization Approach for CNN Model Inference on Integrated GPUs

Leyuan Wang
wangleyu@amazon.com
Amazon Web Services
East Palo Alto, CA, USA

Zhi Chen
chzhi@amazon.com
Amazon Web Services
East Palo Alto, CA, USA

Yizhi Liu
yizhiliu@amazon.com
Amazon Web Services
East Palo Alto, CA, USA

Yao Wang
wayao@amazon.com
Amazon Web Services
East Palo Alto, CA, USA

Lianmin Zheng
lianminzheng@gmail.com
Shanghai Jiaotong University
Shanghai, China

Mu Li
mli@amazon.com
Amazon Web Services
East Palo Alto, CA, USA

Yida Wang
wangyida@amazon.com
Amazon Web Services
East Palo Alto, CA, USA

https://arxiv.org/pdf/1907.02154.pdf

Overall Performance

Models	Ours (ms)	OpenVINO (ms)	Speedup
ResNet50_v1	186.15	203.60	1.09
MobileNet1.0	85.58	53.48	0.62
SqueezeNet1.0	52.10	42.01	0.81
SSD_MobileNet1.0	398.48	–	–
SSD_ResNet50	1006.01	–	–
Yolov3	1004.13	–	–

AWS DeepLens

Models	Ours (ms)	ACL (ms)	Speedup
ResNet50_v1	345.60	358.17	1.04
MobileNet1.0	78.83	95.00	1.21
SqueezeNet1.0	66.61	77.10	1.16
SSD_MobileNet1.0	243.16	216.87	0.89
SSD_ResNet50	777.26	737.90	0.95
Yolov3	1097.47	1042.90	0.95

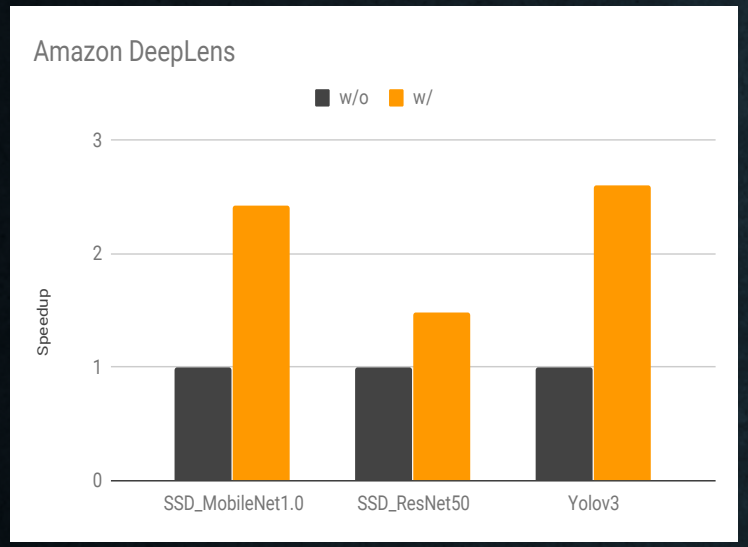
Acer aiSage

Models	Ours (ms)	cuDNN (ms)	Speedup
ResNet50_v1	113.81	117.22	1.03
MobileNet1.0	20.63	30.71	1.49
SqueezeNet1.0	26.58	42.98	1.62
SSD_MobileNet1.0	135.5	197.3	1.47
SSD_ResNet50	371.32	478.33	1.29
Yolov3	553.79	802.41	1.45

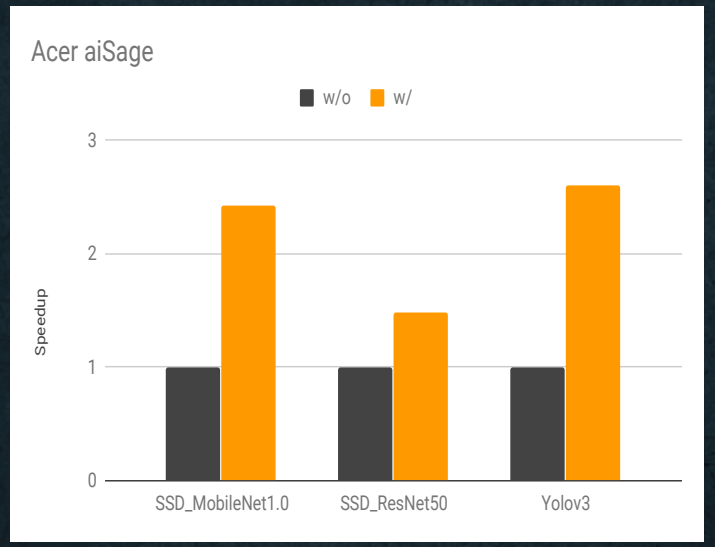
NVIDIA Jetson Nano

aws

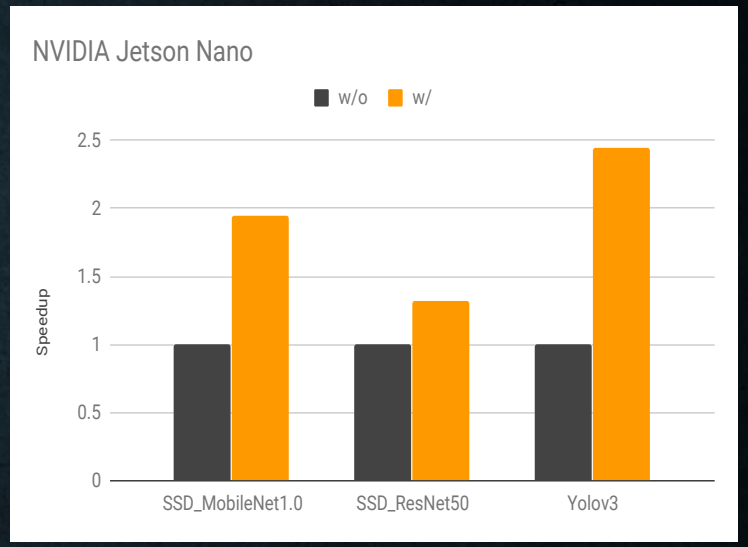
Effects of Vision-specific Optimizations using TVM



AWS DeepLens

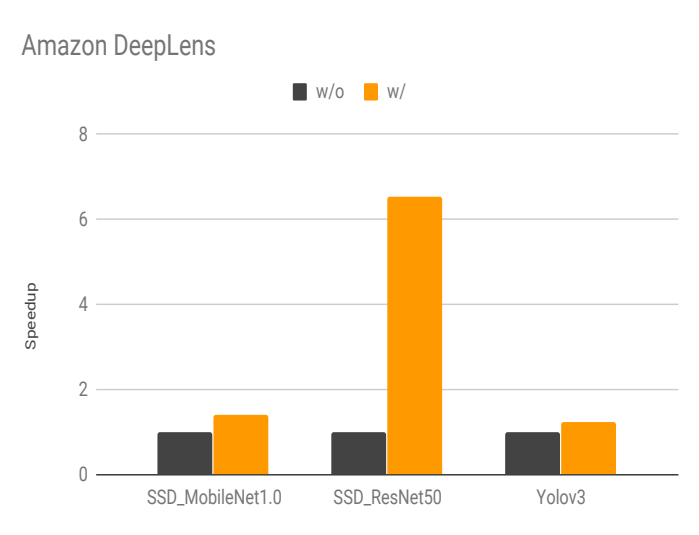


Acer aiSage

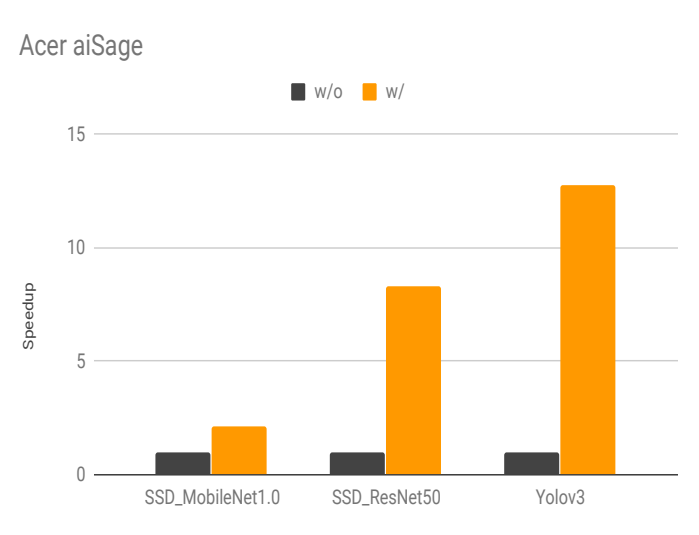


NVIDIA Jetson Nano

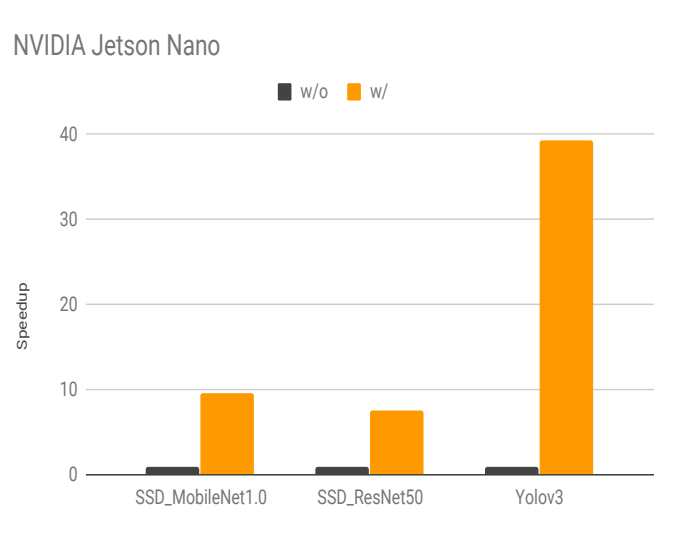
Effects of Convolution operators using TVM



AWS DeepLens



Acer aiSage



NVIDIA Jetson Nano

Like GluonCV? Go build!

https://github.com/dmlc/gluon-cv

https://gluon-cv.mxnet.io

aws

We are hiring!

1. AWS has most TVM contributors from industry.

2. We plan to build TVM team in China,
   based in *Shanghai, Beijing and Shenzhen*.
   1. Applied Scientist and SDE positions
   2. Internship for students interested in ML system.
   3. Research & Development

3. Please contact Yida (wangyida [AT] amazon [DOT] com)
   if interested.