Computer Vision for Embedded Systems

Yung-Hsiang Lu
Purdue University
yunglu@purdue.edu

PURDUE UNIVERSITY

Evaluate Computer Vision

Evaluating Computer Vision

For many people, the only metric is the accuracy using a specific dataset. Even this leaves many questions:

- Which dataset is used?
- Why is this dataset chosen?
- How is accuracy defined?
- What other methods are compared?

Technology Review

Artificial intelligence / Machine learning

Training a single AI model can emit as much carbon as five cars in their lifetimes

Deep learning has a terrible carbon footprint.

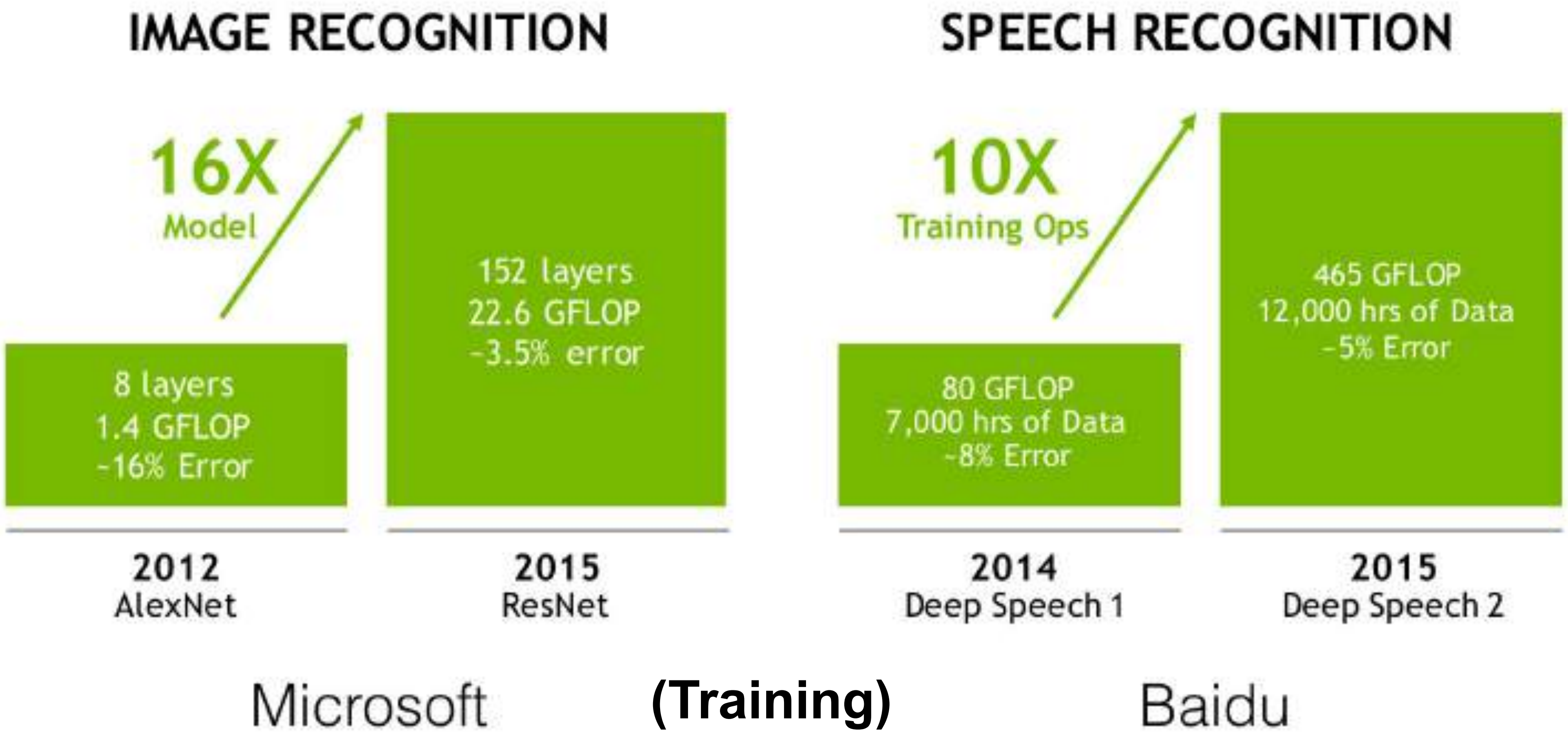
by **Karen Hao**

June 6, 2019

Consumption	CO_2e (lbs)
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000

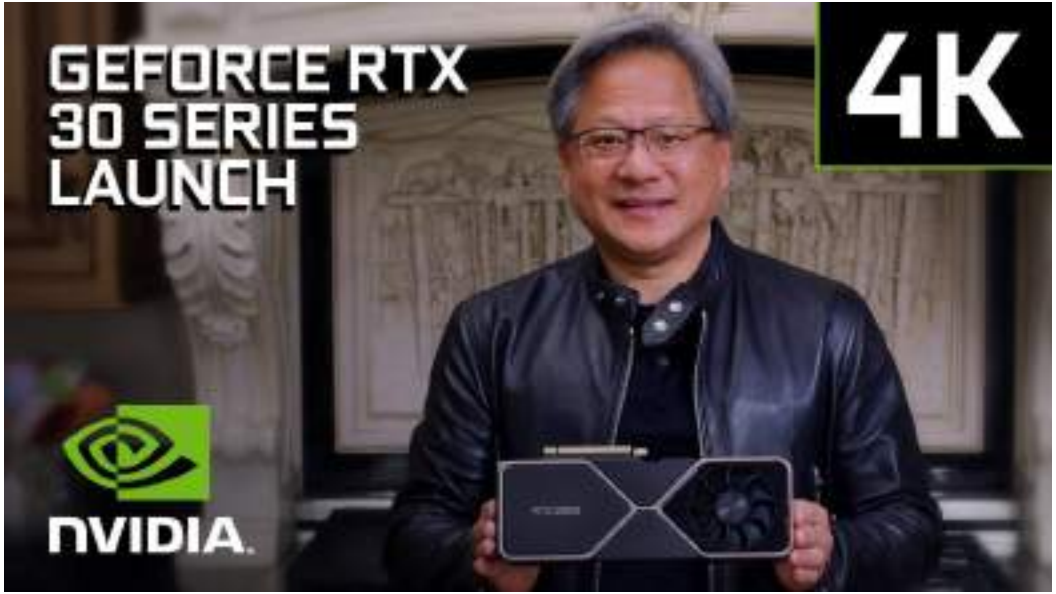
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

Emma Strubell, Ananya Ganesh, Andrew McCallum, "Energy and Policy Considerations for Deep Learning in NLP" 2019



(Training)

Source: cs231n.stanford.edu/slides/2017/cs231n_2017_lecture15.pdf



https://www.newegg.com/gigabyte-geforce-gtx-1080-gv-n1080ix-8gd/p/N82E16814932009

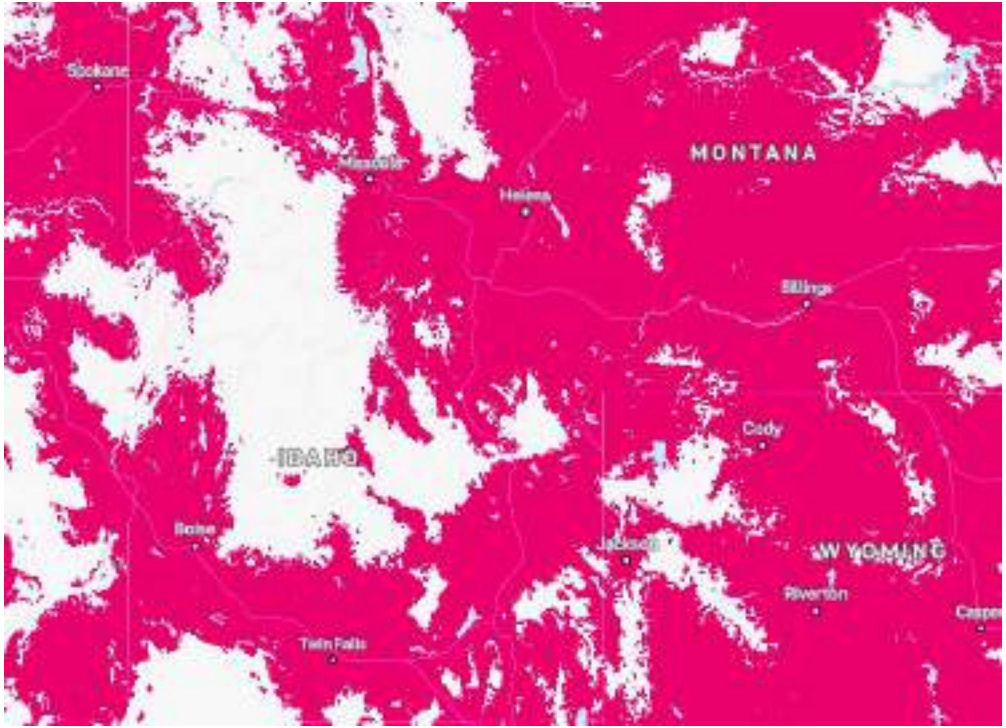
https://www.congatec.com/us/congatec/press-releases/article/congatec-doubles-performance-with-amd-ryzentm-embedded-v2000-processor/

https://www.pcmag.com/picks/the-best-graphics-cards

Transmit all data from cameras to to servers?

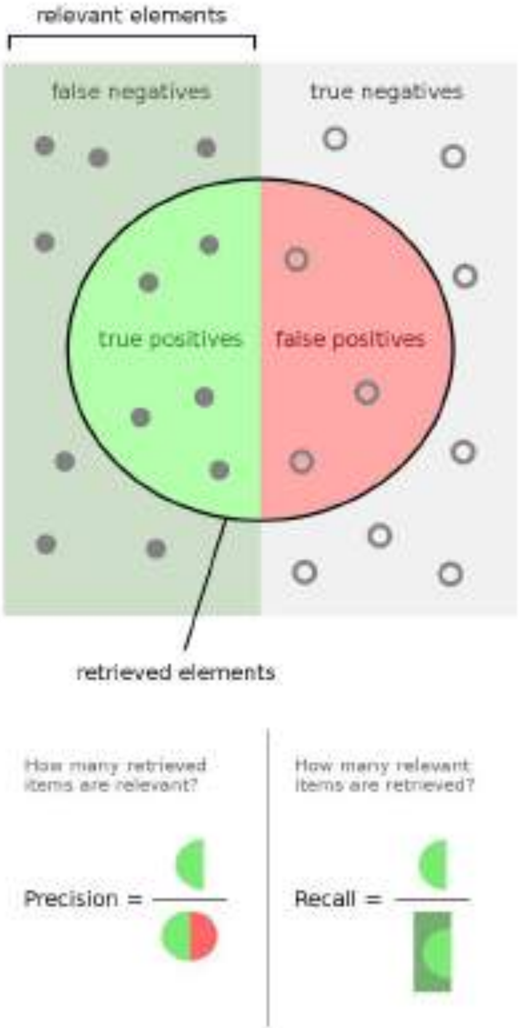
- latency: wireless signals travel at 3.33 microseconds/km
- data rates:
  - Bluetooth up to 3Mb/s, up to 10 meters
  - Wifi (802.11ax) up to 2.4 Gb/s, 70 meters (indoors), 240 outdoors
  - 5G up to 20Gbps, 500 meters
- power: omnidirectional antenna - power proportional to the square of distance. directional antenna can be much more efficient
- privacy: who owns the servers? is data encrypted?
- Homomorphic encryption is not ready yet.

ATT and T Mobile Coverage



Yung-Hsiang Lu, Purdue University

Precision and Recall



source: wikipedia

Factors and metrics for performance

- Accuracy: precision, recall, top-3, top-5, hierarchical
- Execution time: per image (or video frame)
- FPS: frames per second
- FLOPS: floating-point operations
- Memory: to store machine learning model and to process data
- Resolution: number of pixels (width x height)

"Skynet processing at 60 Teraflops a second."
- Movie Terminator 3

Yung-Hsiang Lu, Purdue University

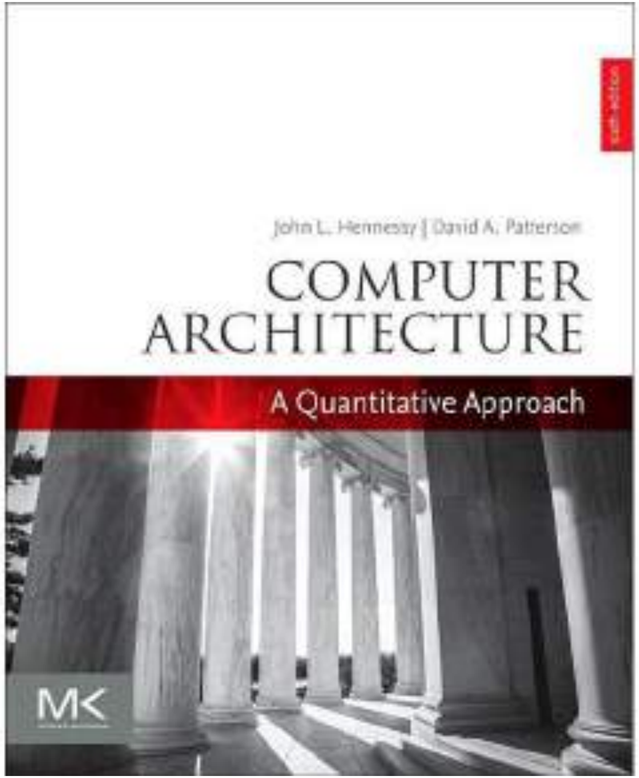
Tuning Parameters (depend on applications)

- Resolution: how many pixels are needed?
- Frame rate: do you really need 30 frames per second?
- Accuracy: estimating crowd or recognizing faces for secure areas?
- General or special purpose?
- Layers of neural networks
- Size of convolution filters

Measuring performance can be complex

GFLOPS/second does not directly translate to performance (execution time)

- Integer operations
- Pipeline processors
- Memory hierarchy
- Thermal throttling
- ...

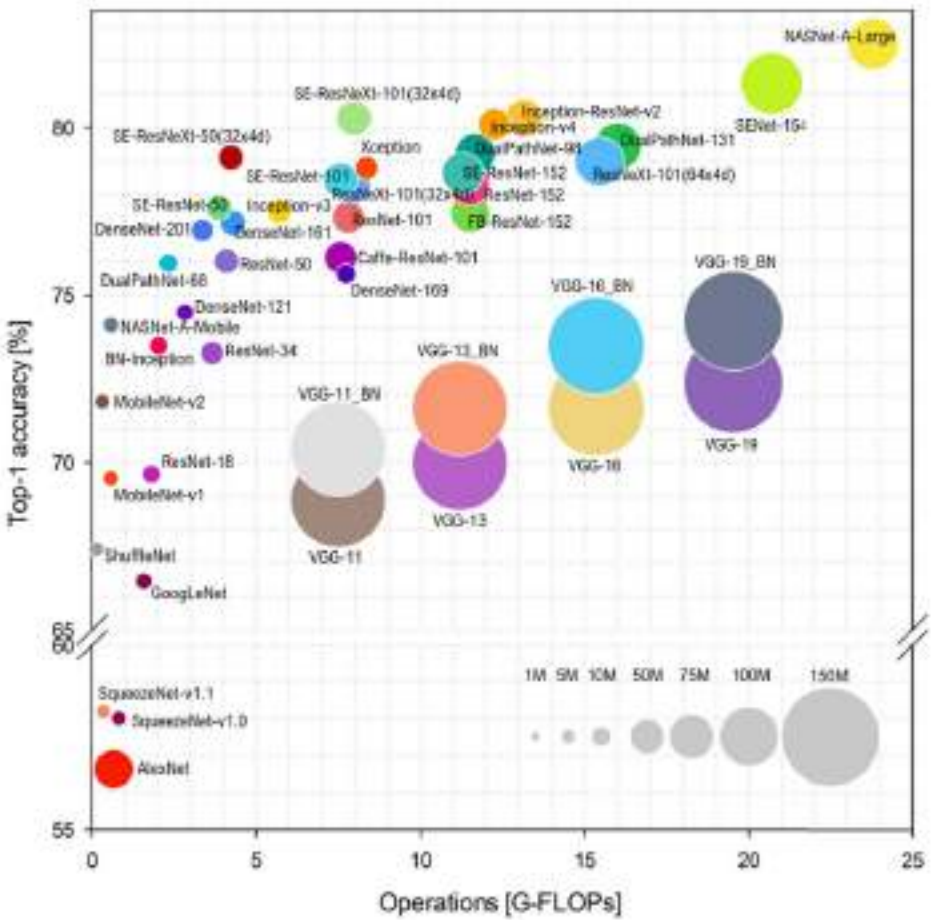


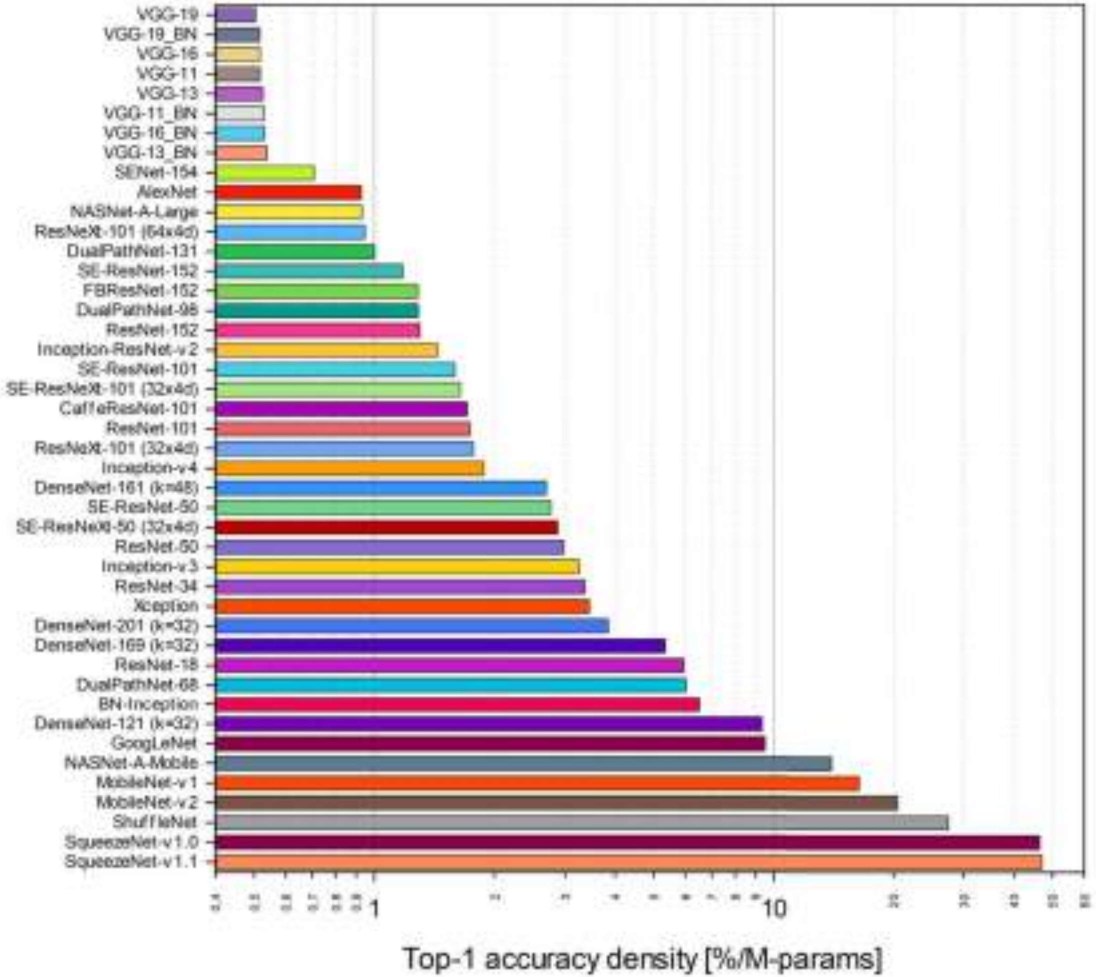
Analysis of Neural Networks

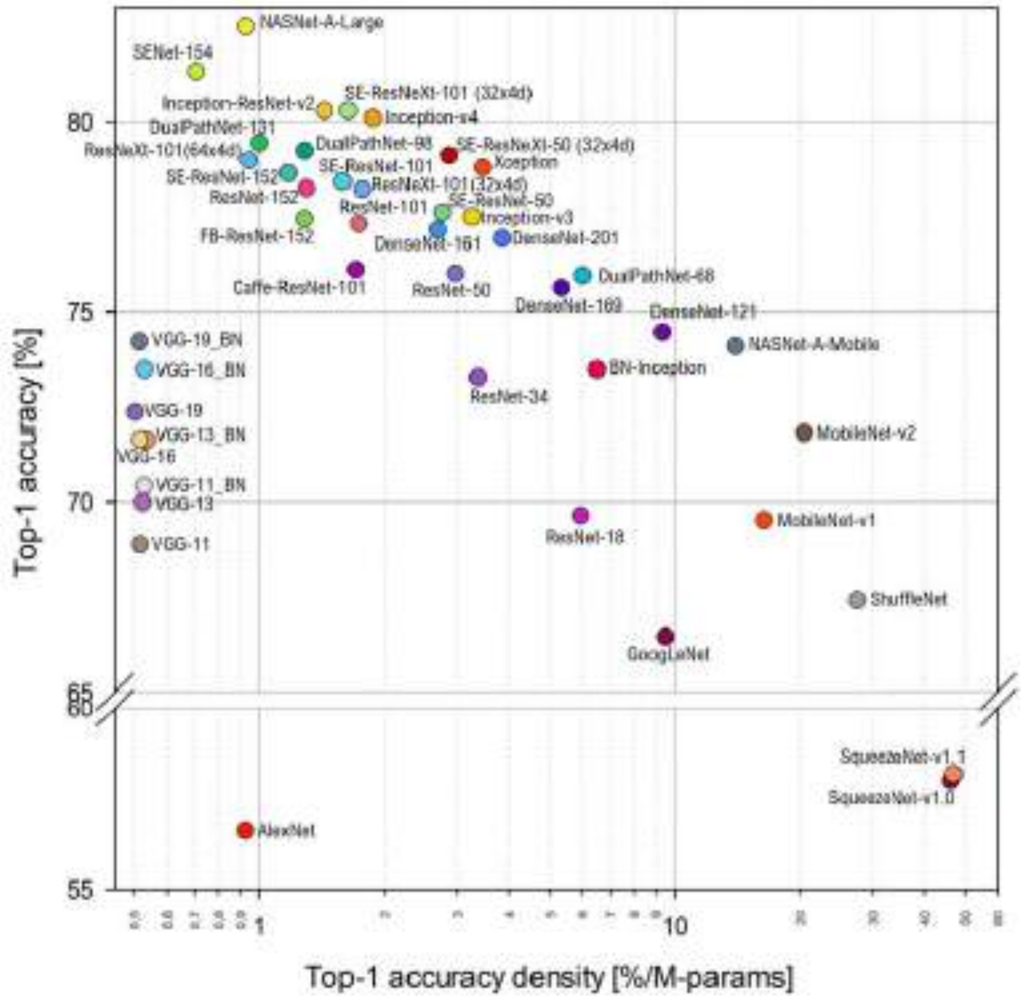
Analysis of Neural Networks

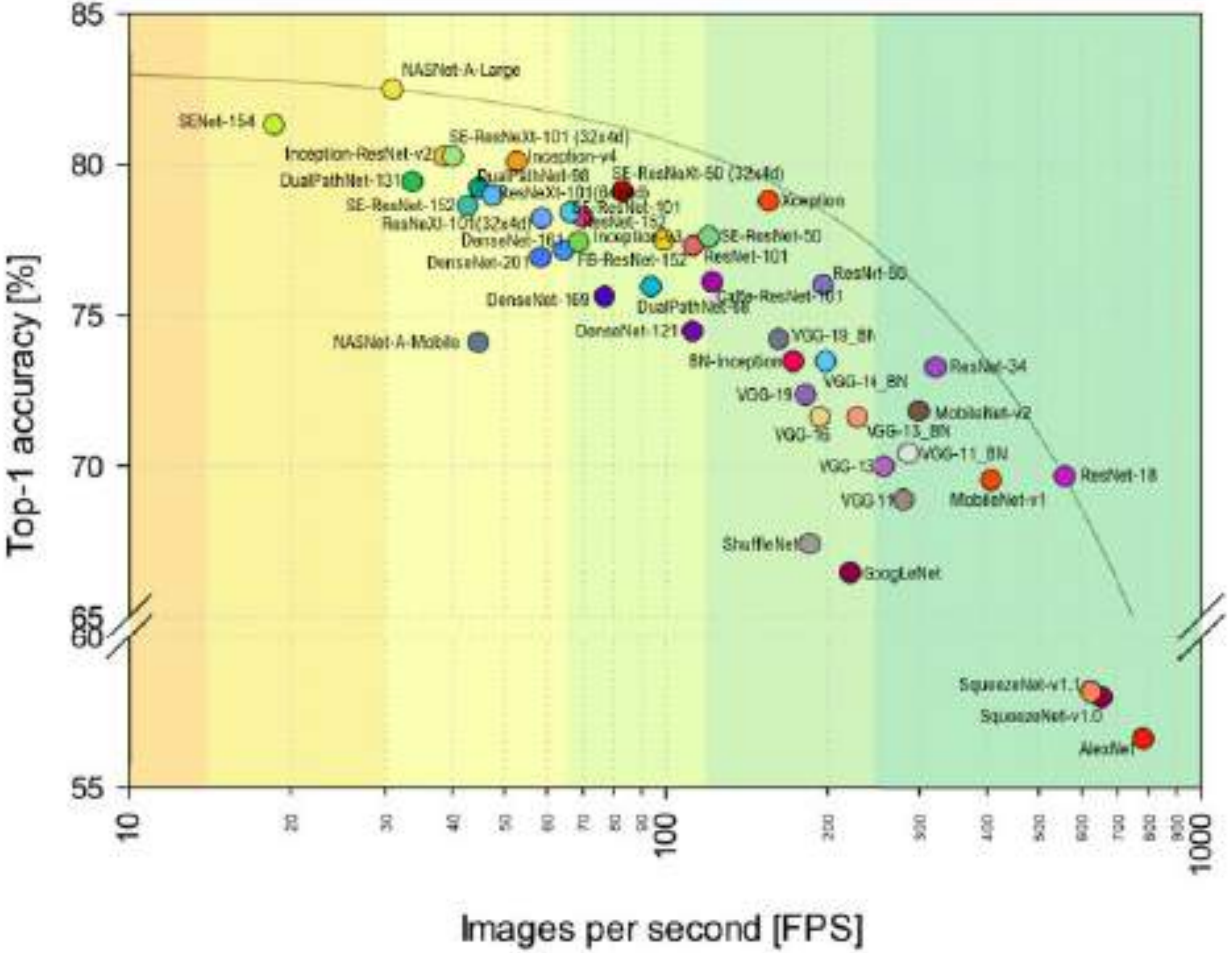
- Which neural networks are analyzed?
- What metrics are used?
- How do these networks perform?
- What patterns can be observed?

Benchmark Analysis of Representative Deep Neural Network Architectures
10.1109/ACCESS.2018.2877890




VGG-19
VGG-19_BN
VGG-16
VGG-11
VGG-13
VGG-11_BN
VGG-16_BN
VGG-13_BN
SENet-154
AlexNet
NASNet-A-Large
ResNeXt-101 (64x4d)
DualPathNet-131
SE-ResNet-152
FBResNet-152
DualPathNet-98
ResNet-152
Inception-ResNet-v2
SE-ResNet-101
SE-ResNeXt-101 (32x4d)
CaffeResNet-101
ResNet-101
ResNeXt-101 (32x4d)
Inception-v4
DenseNet-161 (k=48)
SE-ResNet-50
SE-ResNeXt-50 (32x4d)
ResNet-50
Inception-v3
ResNet-34
Xception
DenseNet-201 (k=32)
DenseNet-169 (k=32)
ResNet-18
DualPathNet-68
BN-Inception
DenseNet-121 (k=32)
GoogLeNet
NASNet-A-Mobile
MobileNet-v1
MobileNet-v2
ShuffleNet
SqueezeNet-v1.0
SqueezeNet-v1.1
1
10
Top-1 accuracy density [%/M-params]


NASNet-A-Large
SENet-154
SE-ResNeXt-101 (32x4d)
Inception-ResNet-v2
Inception-v4
DualPathNet-131
DualPathNet-98
SE-ResNeXt-50 (32x4d)
ResNeXt-101 (64x4d)
Xception
SE-ResNet-152
SE-ResNet-101
ResNeXt-101 (32x4d)
ResNet-152
SE-ResNet-50
ResNet-101
Inception-v3
FB-ResNet-152
DenseNet-161
DenseNet-201
Caffe-ResNet-101
ResNet-50
DualPathNet-68
DenseNet-169
DenseNet-121
VGG-19_BN
NASNet-A-Mobile
VGG-16_BN
BN-Inception
ResNet-34
VGG-19
VGG-13_BN
MobileNet-v2
VGG-16
VGG-11_BN
VGG-13
ResNet-18
MobileNet-v1
VGG-11
ShuffleNet
GoogLeNet
SqueezeNet-v1.1
SqueezeNet-v1.0
AlexNet
80
75
70
65
60
55
1
10
Top-1 accuracy [%]
Top-1 accuracy density [%/M-params]


Top-1 accuracy [%]
Images per second [FPS]
NASNet-A-Large
SENet-154
SE-ResNeXt-101 (32x4d)
Inception-ResNet-v2
Inception-v4
DualPathNet-131
DualPathNet-98
SE-ResNeXt-50 (32x4d)
SE-ResNet-152
Xception
SE-ResNet-101
ResNet-152
DenseNet-161
SE-ResNet-50
DenseNet-201
FB-ResNet-152
ResNet-101
ResNet-50
DenseNet-169
Caffe-ResNet-101
DenseNet-121
NASNet-A-Mobile
BN-Inception
ResNet-34
VGG-19
MobileNet-v2
VGG-16
VGG-13_BN
VGG-11_BN
VGG-13
ResNet-18
MobileNet-v1
ShuffleNet
GoogLeNet
SqueezeNet-v1.1
SqueezeNet-v1.0
AlexNet

Compare Networks by Training

Application	Model	Number of Layers	Dominant Layer	Implementations	Dataset
Image classification	ResNet-50 [56] Inception-v3 [80]	50 (152 max) 42	CONV	TensorFlow, MXNet, CNTK	ImageNet1K [73]
Machine translation	Seq2Seq [79] Transformer [82]	5 12	LSTM Attention	TensorFlow, MXNet TensorFlow	IWSLT15 [21] WMT-14 [18]
Object detection	Faster R-CNN [71]	101[a]	CONV	TensorFlow, MXNet	Pascal VOC 2007 [37]
Speech recognition	Deep Speech 2 [13]	9[b]	RNN	MXNet	LibriSpeech [64]
Adversarial learning	WGAN [40]	14+14[c]	CONV	TensorFlow	Downsampled ImageNet [29]
Deep reinforcement learning	A3C [62]	4	CONV	MXNet	Atari 2600

Dataset	Number of Samples	Size	Special
ImageNet1K	1.2million	3x256x256 per image	N/A
IWSLT15	133k	20-30 words long per sentence	vocabulary size of 17188 (English to Vietnamese)
WMT-14	4.5million	up to 50 words (most sentences)	vocabulary size of 37000 (English to German)
Pascal VOC 2007	5011[d]	around 500x350	12608 annotated objects
LibriSpeech	280k	1000 hours[e]	N/A
Downsampled ImageNet	1.2million	3x64x64 per image	N/A
Atari 2600	N/A	4x84x84 per image	N/A

H. Zhu et al., "Benchmarking and Analyzing Deep Neural Network Training," 2018 IEEE International Symposium on Workload Characterization (IISWC), 2018, pp. 88-100, doi: 10.1109/IISWC.2018.8573476.

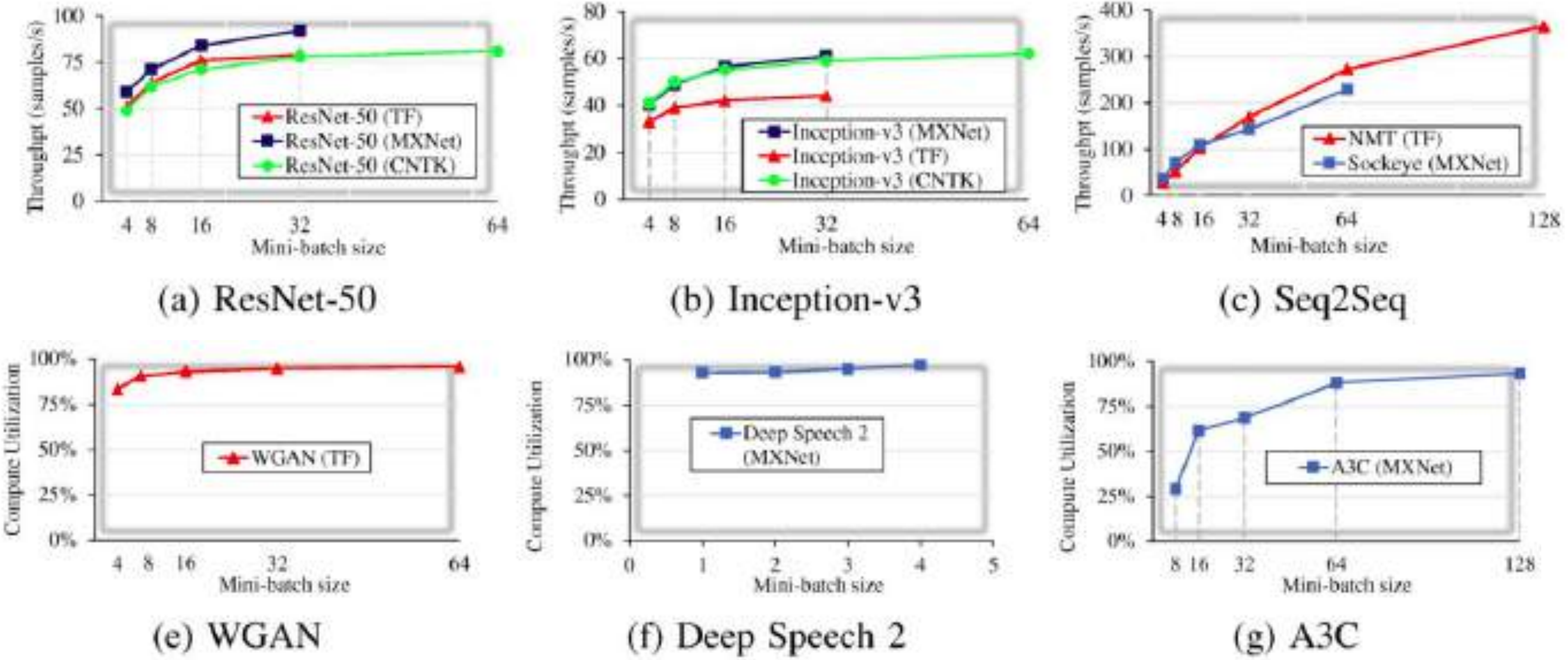
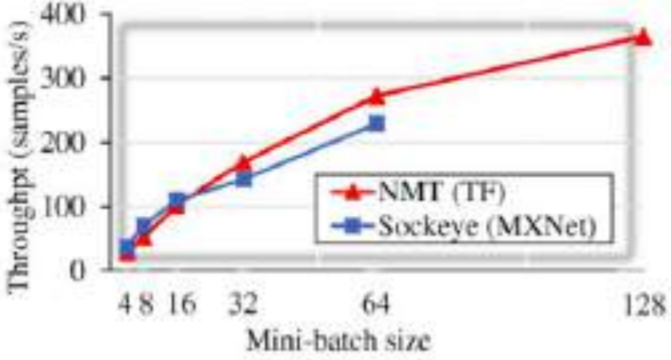
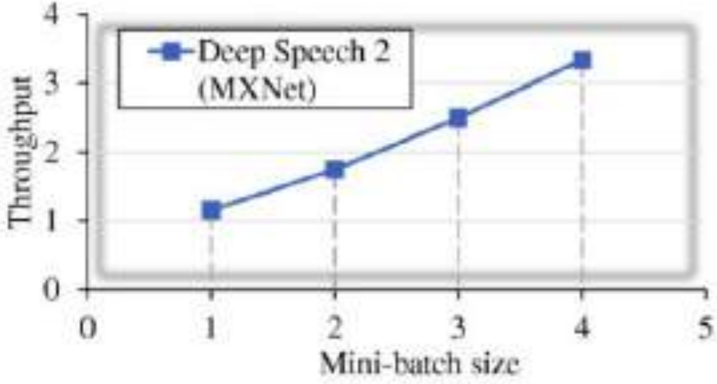


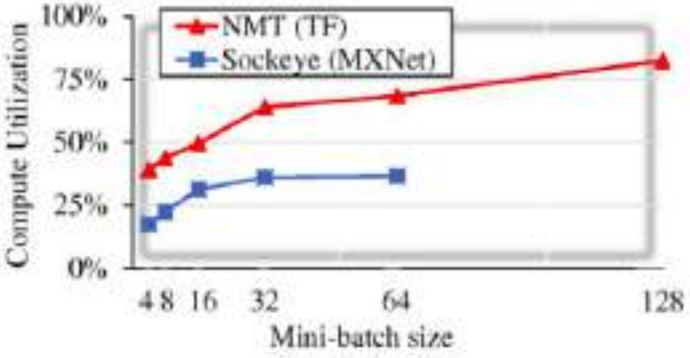
Fig. 5: GPU compute utilization for different models on multiple mini-batch sizes.



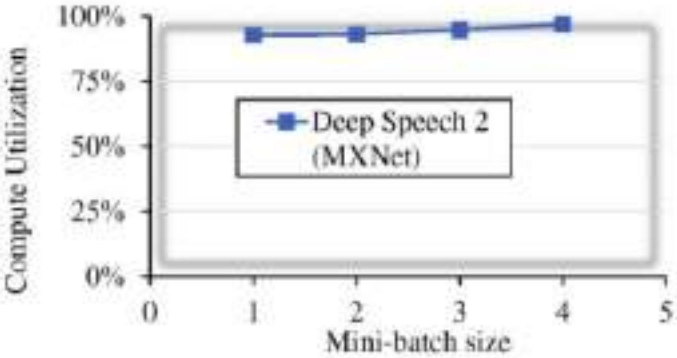
(c) Seq2Seq



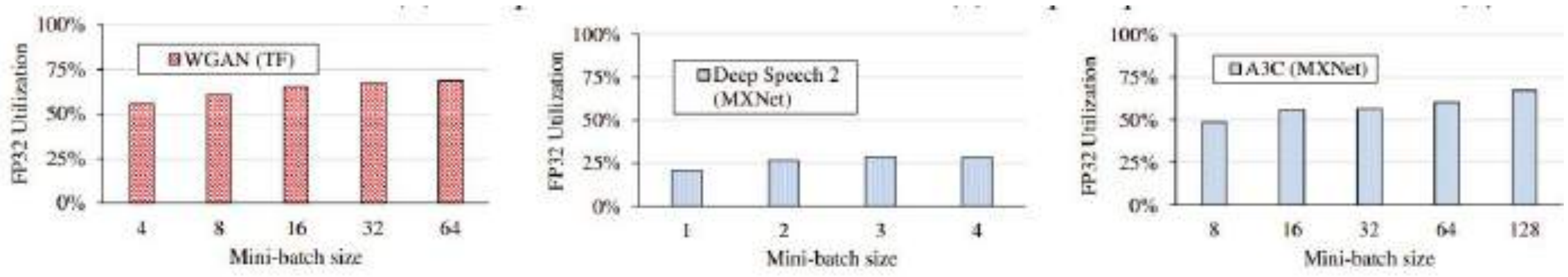
(f) Deep Speech 2



(c) Seq2Seq

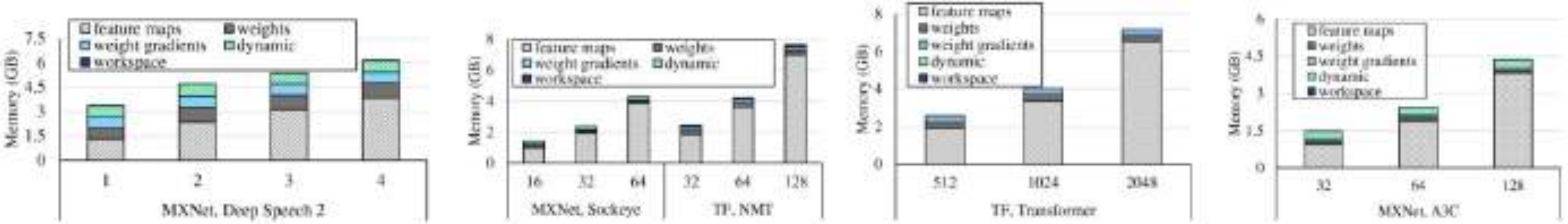


(f) Deep Speech 2



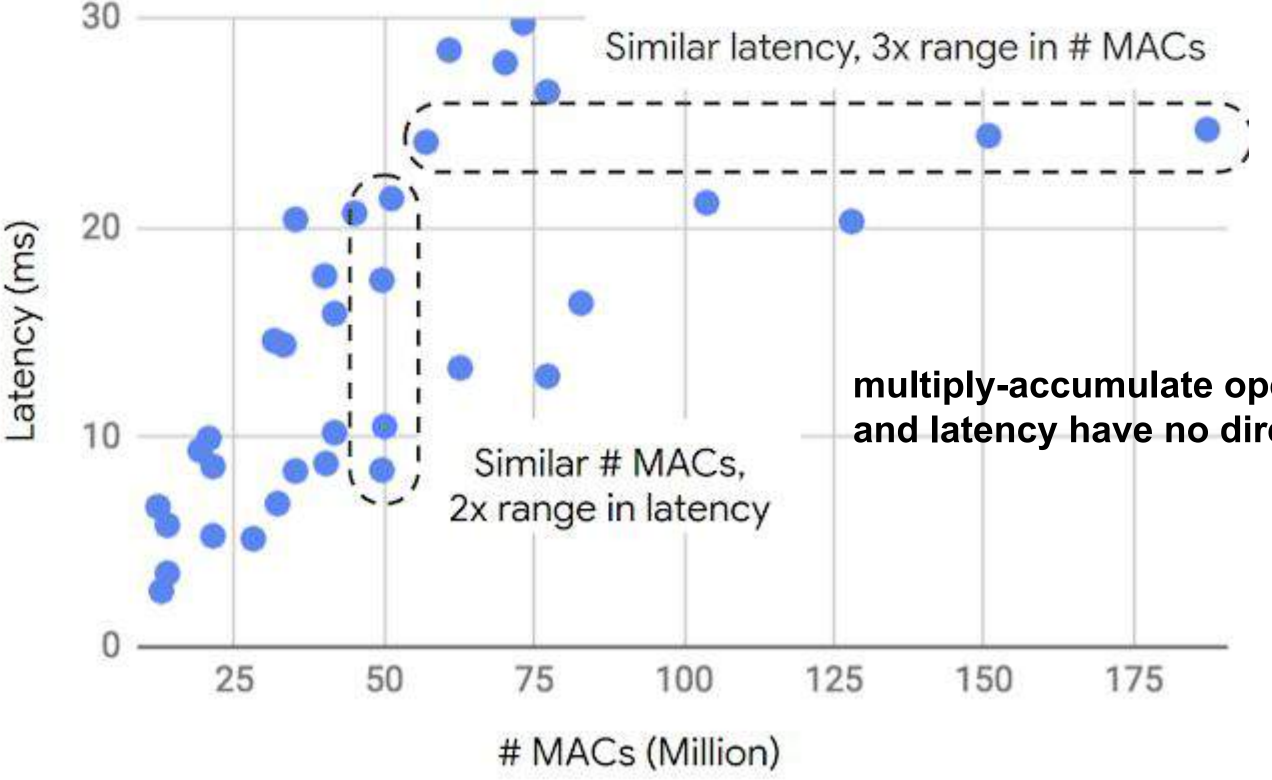
(e) WGAN (f) Deep Speech 2 (g) A3C

Fig. 6: GPU FP32 utilization for different models on multiple mini-batch sizes.



(d) Deep Speech 2 (e) Seq2Seq (f) Transformer (g) A3C

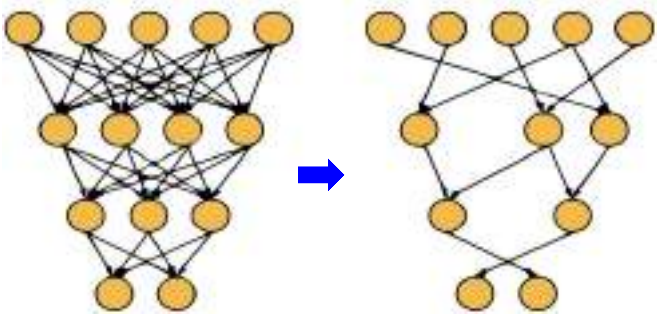
Fig. 8: GPU memory usage breakdown for different models on multiple mini-batch sizes.



multiply-accumulate operations (MACs) and latency have no direct relationship

Introducing the CVPR 2018 On-Device Visual Intelligence Challenge
Friday, April 20, 2018, Google AI Blog

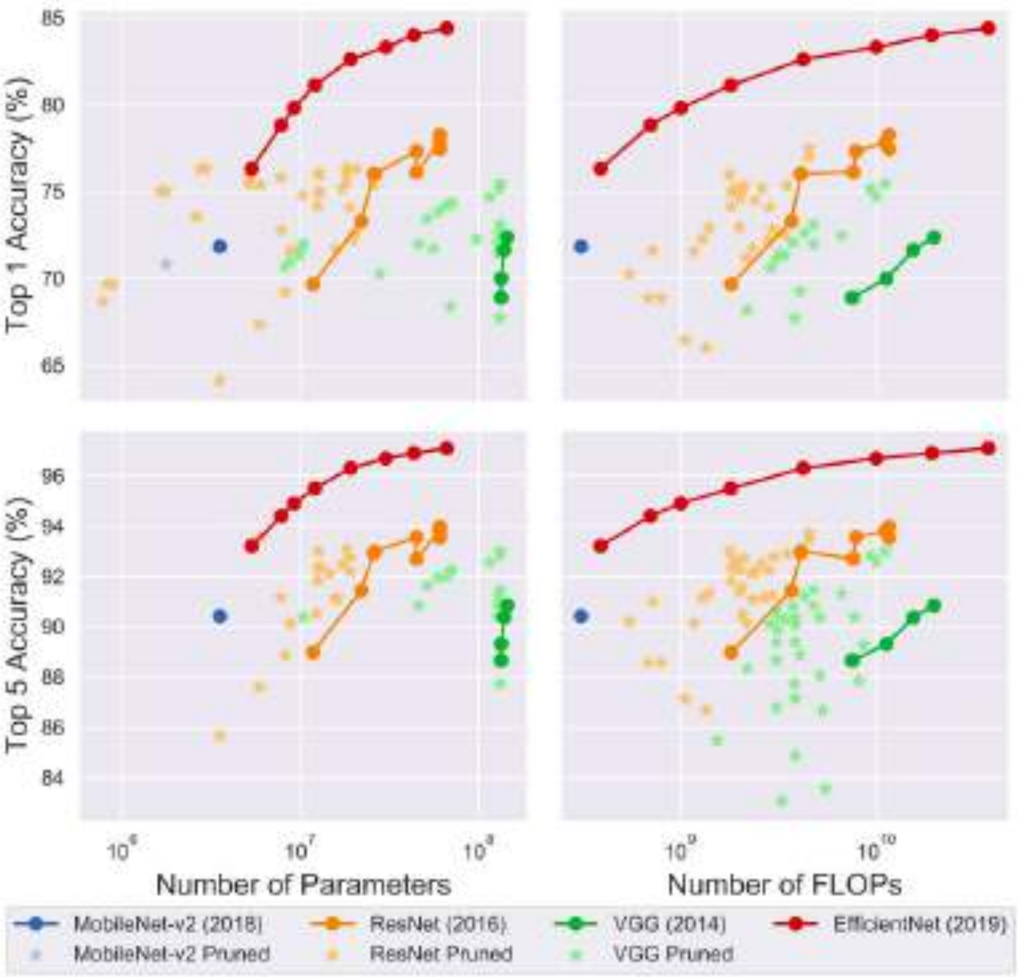
Network Pruning

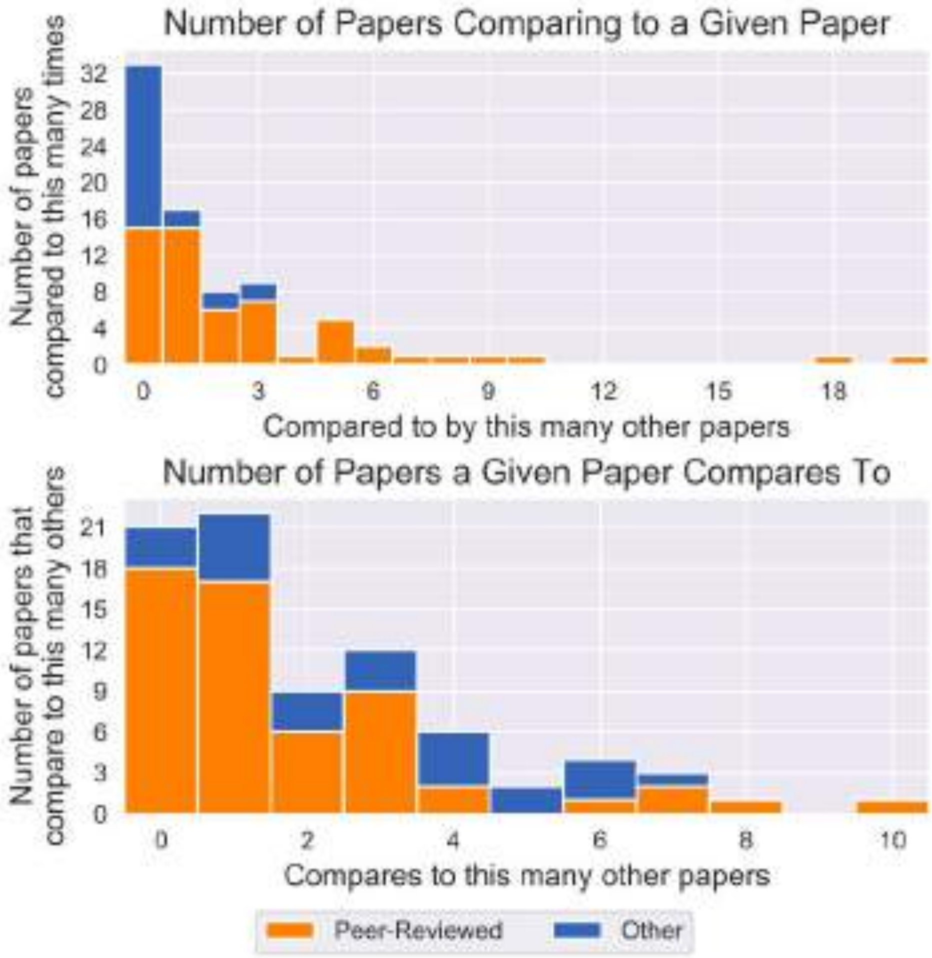


$$f\left(\sum_{k=1}^{n} w_k \times i_k + b\right)$$

What is the state of neural network pruning?

Davis Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle, John Guttag




Number of Papers Comparing to a Given Paper
Number of papers compared to this many times
Compared to by this many other papers
Number of Papers a Given Paper Compares To
Number of papers that compare to this many others
Compares to this many other papers
Peer-Reviewed
Other

Reproducibility Challenge

Reproducing results in research papers can be hard:

- lack of source code
- comparable hardware
- software environment, library and right versions
- undocumented parameters

Summary

- Computer vision can be evaluated in many different ways, including performance.
- Performance can be defined in different ways, such as execution time.
- Many factors affect performance, such as the sizes of the networks, but the relationships are not straight lines.
- Training time is affected by the sizes of mini batches.