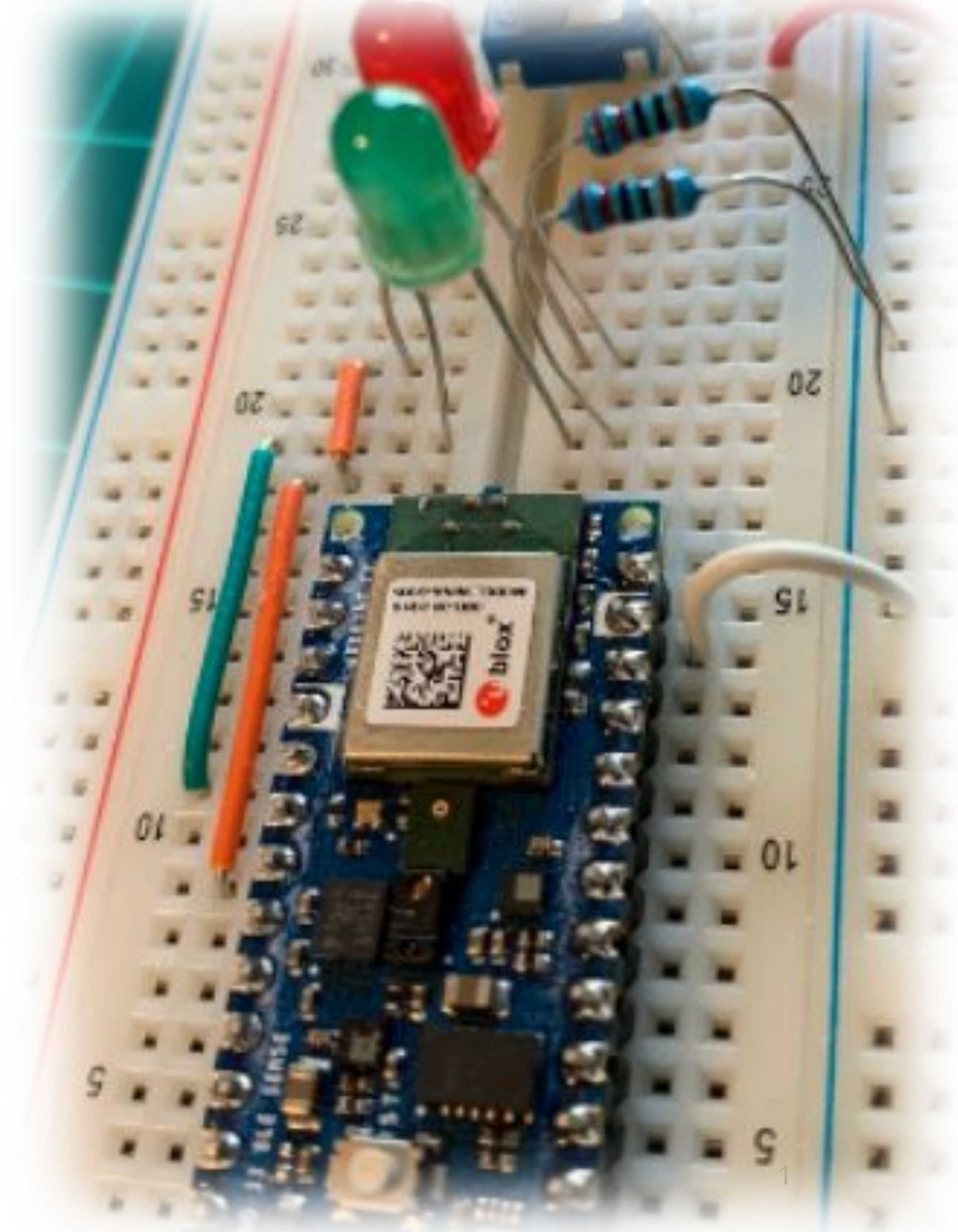
IESTI01 – TinyML

Embedded Machine Learning

26. Visual Wake Words (VWW) Introduction

Prof. Marcelo Rovai
UNIFEI



What is VWW,
Visual Wake Words?

Person Detection

Mask Detection

Person Detection

Next **big** thing!

Augmented Reality

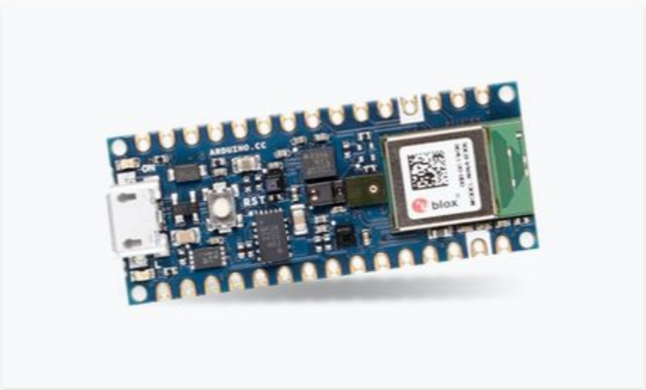
- Smart Shopping

Next **big** thing!

Augmented Reality

- Smart Shopping
- Navigation

Opportunities



Run visual wake words
on-device

Visual Wake Words, Challenges

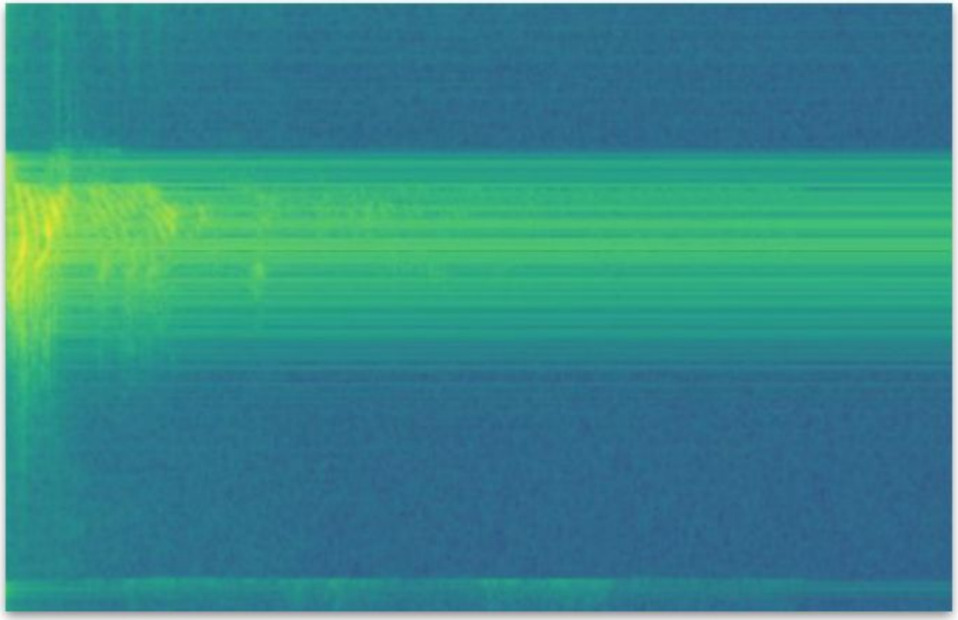
Simple **Experiment**

$$\underbrace{224 \times 224}_{\text{Pixels}} \times \underset{\substack{\text{RGB} \\ \text{(\# channels)}}}{3} \times \underset{\text{Bytes/Pixel}}{4} = 602{,}112 \text{ Bytes}$$

224

224

40



49

$$49 \times 40 \times 1 \times 4 = 7{,}840 \text{ Bytes}$$

Pixels

RGB
(# channels)

Bytes/Pixel

$$224 \times 224 \times 3 \times 4 = 602{,}112 \text{ Bytes}$$

Pixels

RGB
(# channels)

Bytes/Pixel

224

224

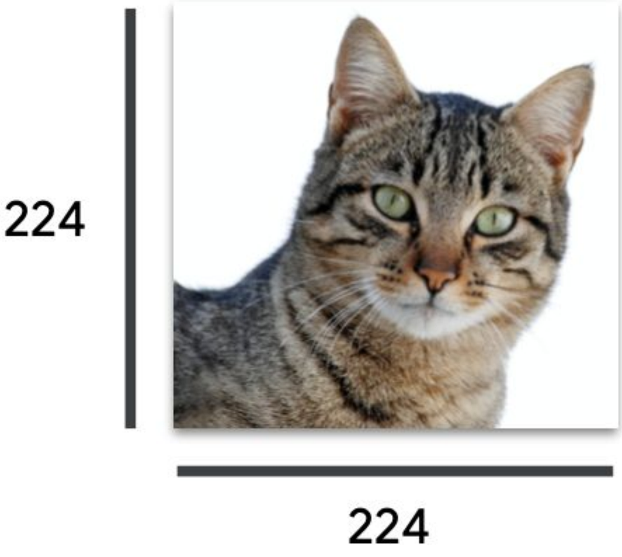
Simple **Experiment**

Always-on (Visual Wake Words)?

→ Much more data (than KWS)

- Higher **latency**
- Higher **power consumption** (drains battery)

  → Lower **user satisfaction**

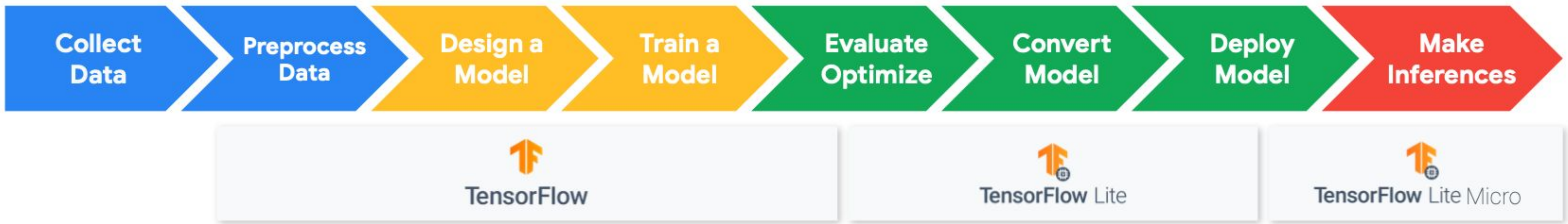


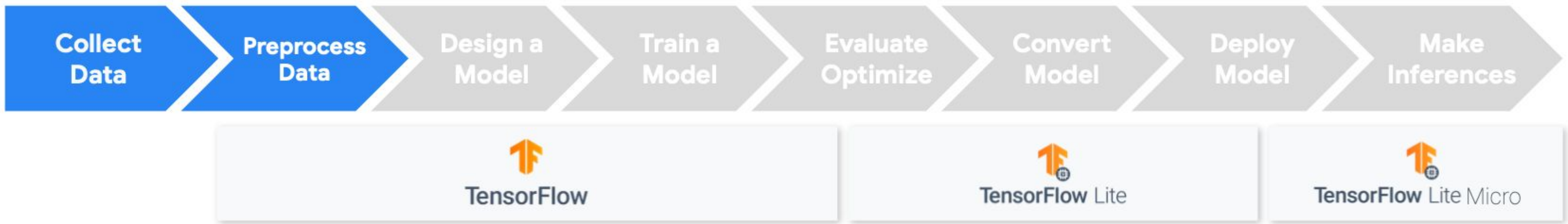
Memory (CNN Models)

Model	Size	Top-1 Accuracy
Xception	88 MB	0.790
VGG16	528 MB	0.713
ResNet50	98 MB	0.749
Inception v3	92 MB	0.779
MobileNet v1	16 MB	0.713
DenseNet 201	80 MB	0.773
NASNetMobile	23 MB	0.825

Our board
has **256 KB** of RAM (memory)

Visual Wake Words,
Data Collection and Processing


Collect Data
Preprocess Data
Design a Model
Train a Model
Evaluate Optimize
Convert Model
Deploy Model
Make Inferences
TensorFlow
TensorFlow Lite
TensorFlow Lite Micro


Collect Data
Preprocess Data
Design a Model
Train a Model
Evaluate Optimize
Convert Model
Deploy Model
Make Inferences
TensorFlow
TensorFlow Lite
TensorFlow Lite Micro

Visual Wake Words **Dataset**

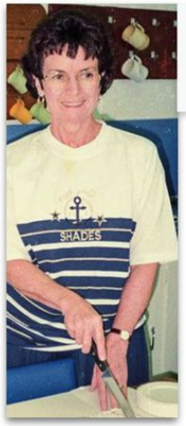
Visual Wake Words Dataset

Aakanksha Chowdhery, Pete Warden, Jonathon Shlens,
Andrew Howard, Rocky Rhodes
Google Research
{chowdhery, petewarden, shlens, howarda, rocky}@google.com

https://arxiv.org/pdf/1906.05721.pdf

Visual Wake Words **Dataset**

Label: **“person”**



Label: **“person”**

Label: **“not-person”**

(Labeled from COCO dataset)

Visual Wake Words **Dataset**

Data collection is **DIFFICULT**

- This dataset and collection process is ***limited*** and has bias
- Small number of relevant images
- Large quantity of irrelevant images

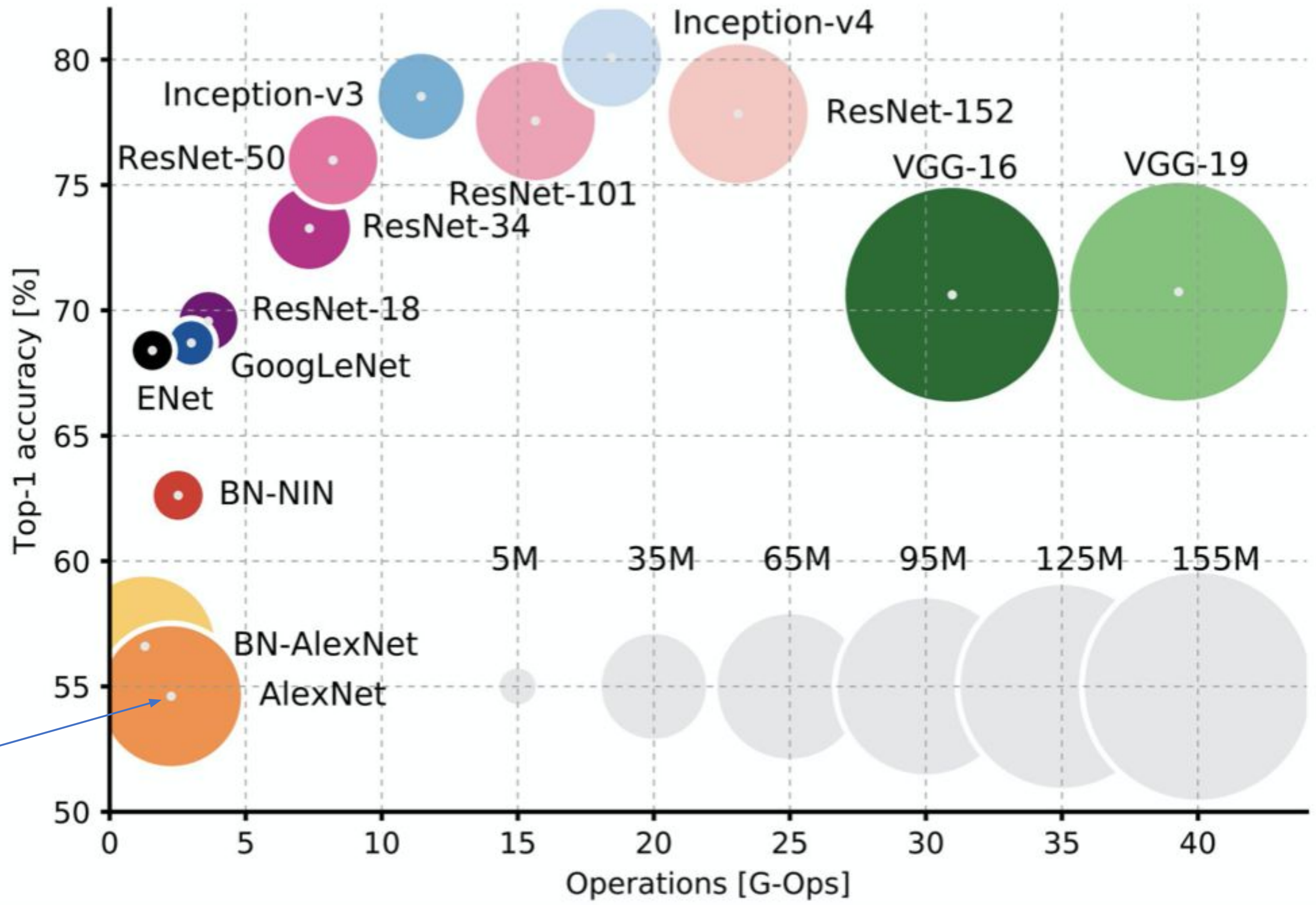
Visual Wake Words Dataset

Aakanksha Chowdhery, Pete Warden, Jonathon Shlens,
Andrew Howard, Rocky Rhodes
Google Research
{chowdhery, petewarden, shlens, howarda, rocky}@google.com

Visual Wake Words, Model


Collect Data
Preprocess Data
Design a Model
Train a Model
Evaluate Optimize
Convert Model
Deploy Model
Make Inferences
TensorFlow
TensorFlow Lite
TensorFlow Lite Micro

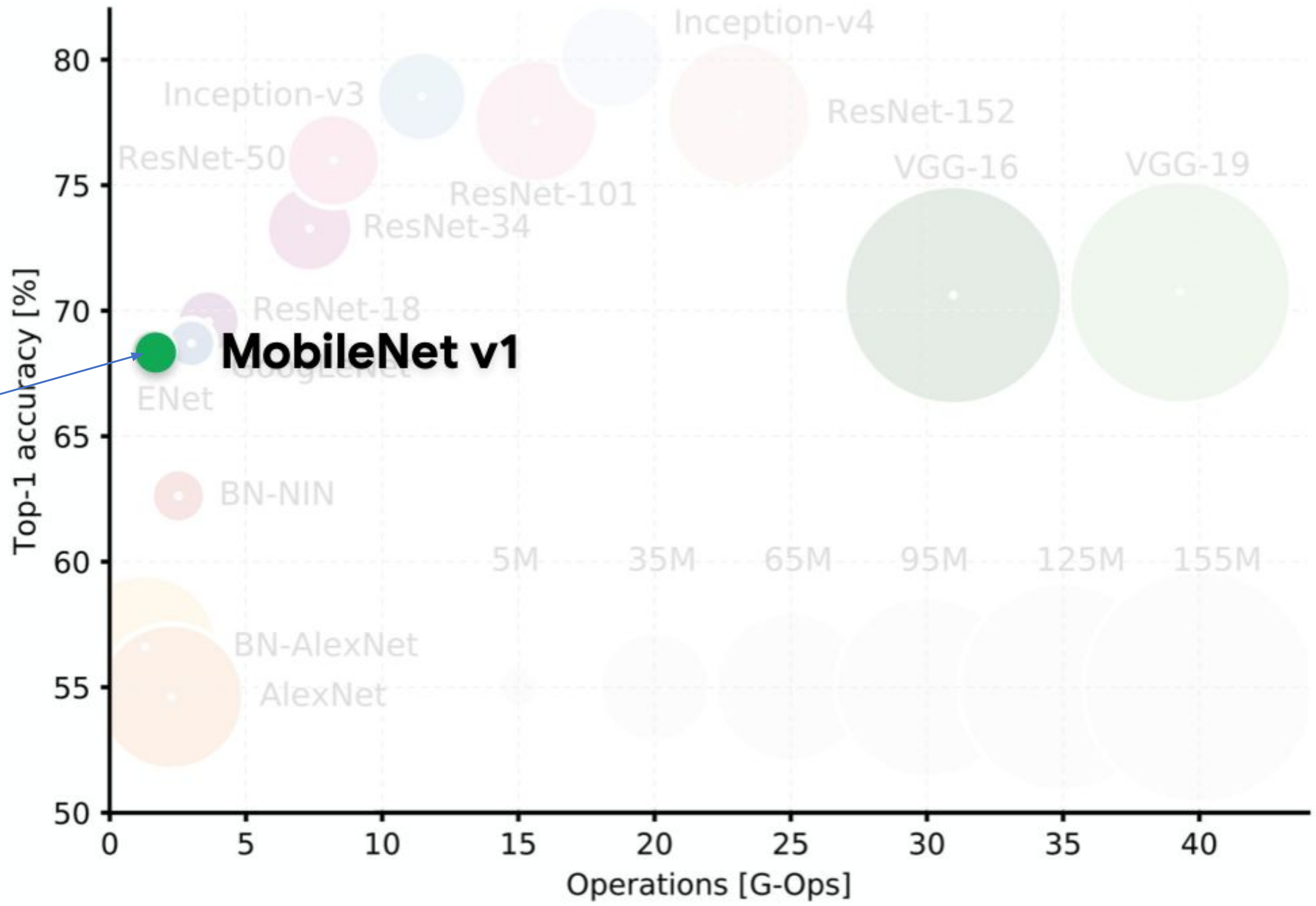
Model **Evolution**



(2012)

Model **Evolution**

(2017)



MobileNet v1

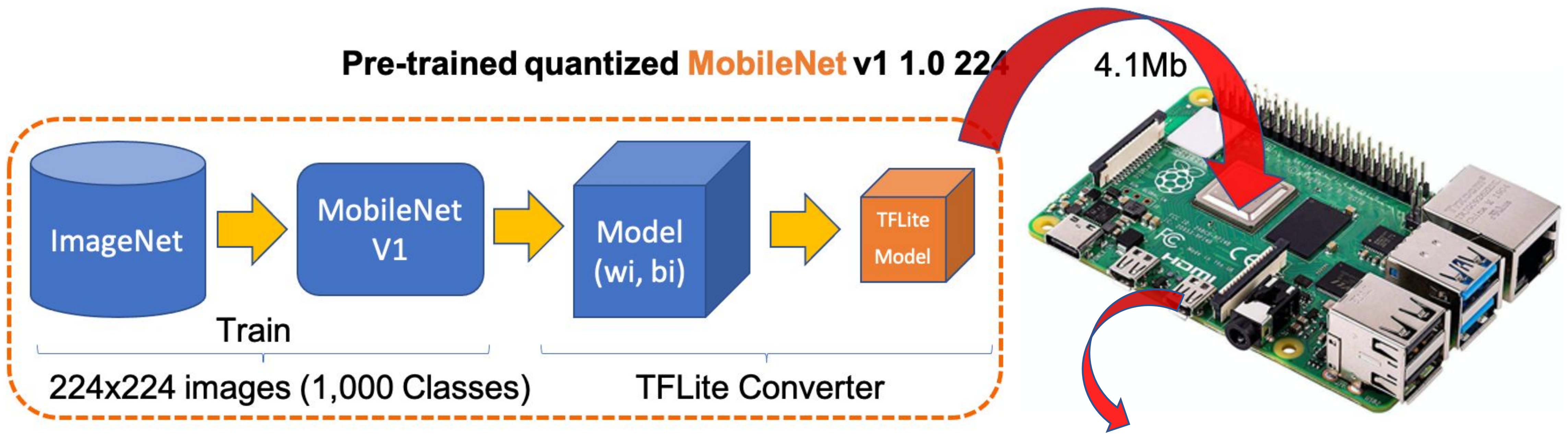
MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications

Andrew G. Howard Menglong Zhu Bo Chen Dmitry Kalenichenko
Weijun Wang Tobias Weyand Marco Andreetto Hartwig Adam

Google Inc.

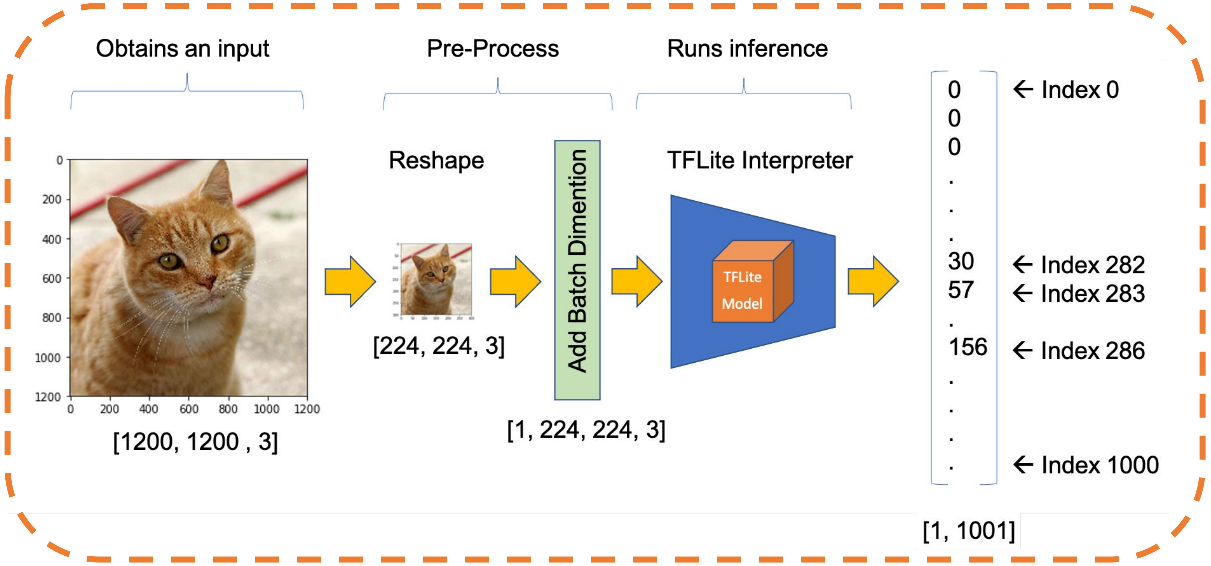
{howarda,menglong,bochen,dkalenichenko,weijunw,weyand,anm,hadam}@google.com

https://arxiv.org/pdf/1704.04861.pdf



[Exploring IA at the Edge!]

[EdgeML with TF-Lite - RPi)]



MobileNet v1

Model	Size	Top-1 Accuracy
MobileNet v1	***16 MB*** *	0.713

* Not Quantized

Fine for mobile phones with GB of RAM, but 64X microcontroller RAM

Our board [Course 3 Kit] only has **256KB** of RAM (memory)

Further **Optimizations**

Multiply-Accumulates

α	Image Size	MACs (millions)	Params (millions)	Top-1 Accuracy
1	224	569	**4.24**	**70.7**
1	128	186	4.14	64.1
0.75	224	317	**2.59**	**68.4**
0.75	128	104	2.59	61.8
0.5	224	150	**1.34**	**64.0**
0.5	128	49	1.34	56.2
0.25	224	41	**0.47**	**50.6**
0.25	128	14	0.47	41.2

Model

MobileNetV1 96x96 0.25

A pre-trained multi-layer convolutional network designed to efficiently classify images. Uses around 105.9K RAM and 301.6K ROM with default settings and optimizations. Works best with 96x96 input size. Supports both RGB and grayscale.

Image Size

MobileNetV1 96x96 0.2

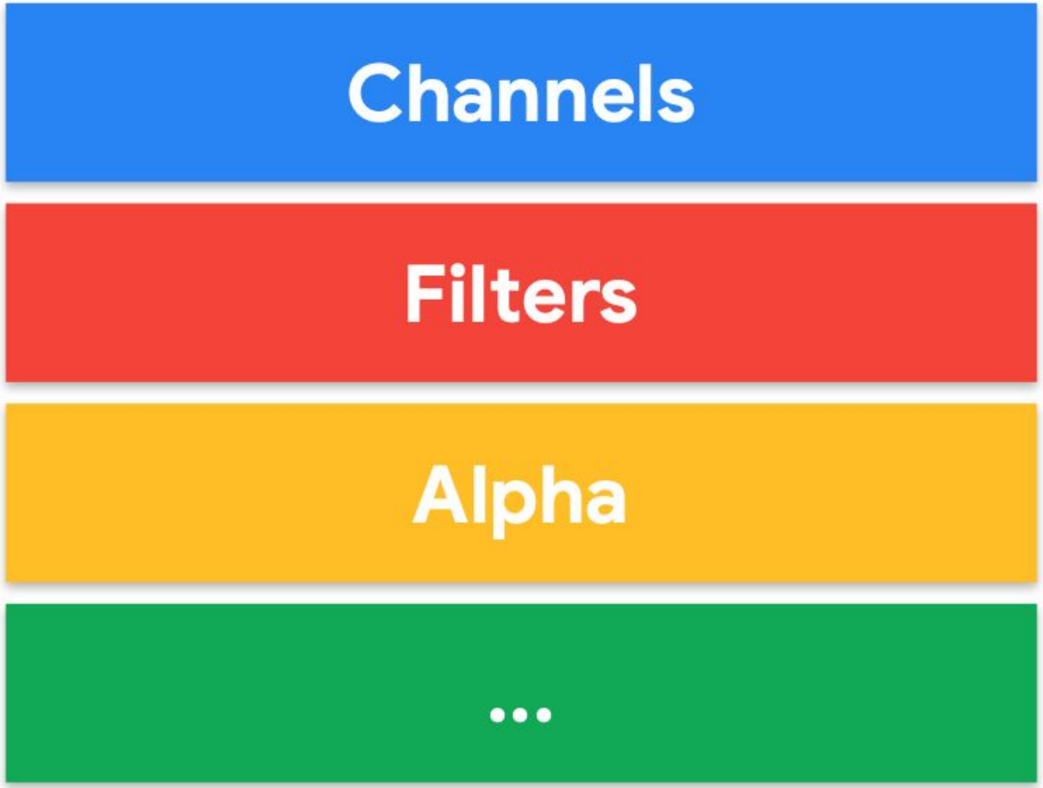
Uses around 83.1K RAM and 218.3K ROM with default settings and optimizations. Works best with 96x96 input size. Supports both RGB and grayscale.

Alpha

MobileNetV1 96x96 0.1

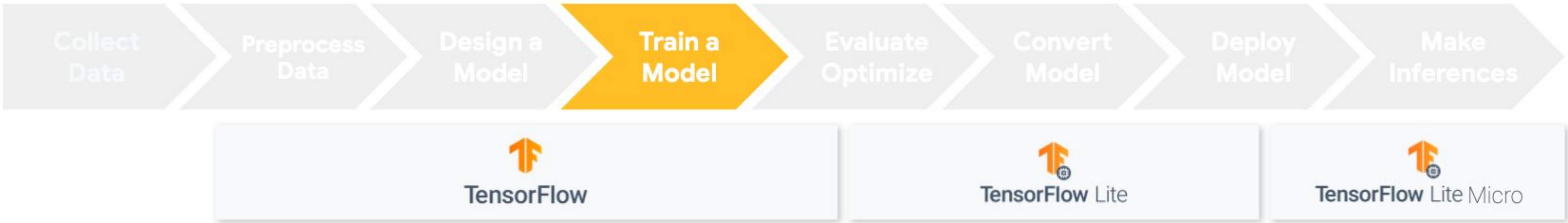
Uses around 53.2K RAM and 101K ROM with default settings and optimizations. Works best with 96x96 input size. Supports both RGB and grayscale.

Neural Architecture Search (NAS)

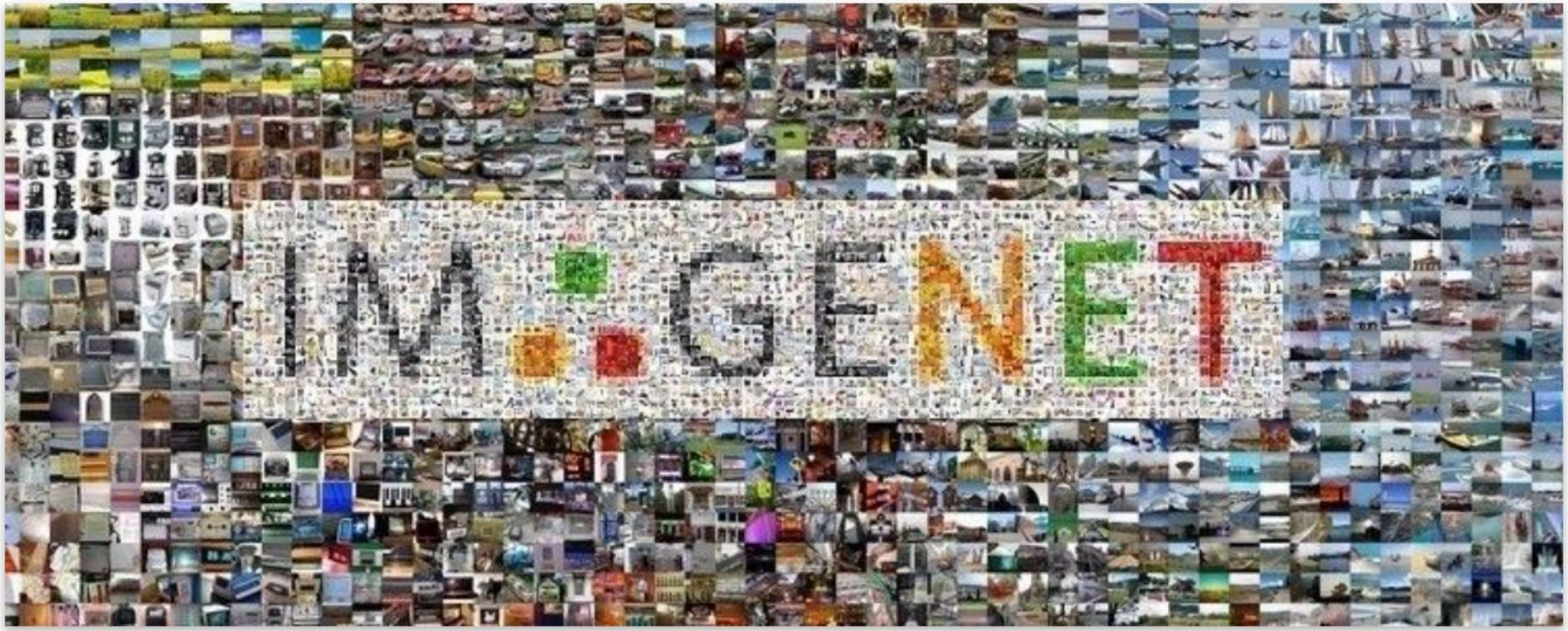


https://towardsdatascience.com/what-is-neural-architecture-search-and-why-should-you-care-1e22393de461

Visual Wake Words,
Training a Model


Collect Data
Preprocess Data
Design a Model
Train a Model
Evaluate Optimize
Convert Model
Deploy Model
Make Inferences
TensorFlow
TensorFlow Lite
TensorFlow Lite Micro

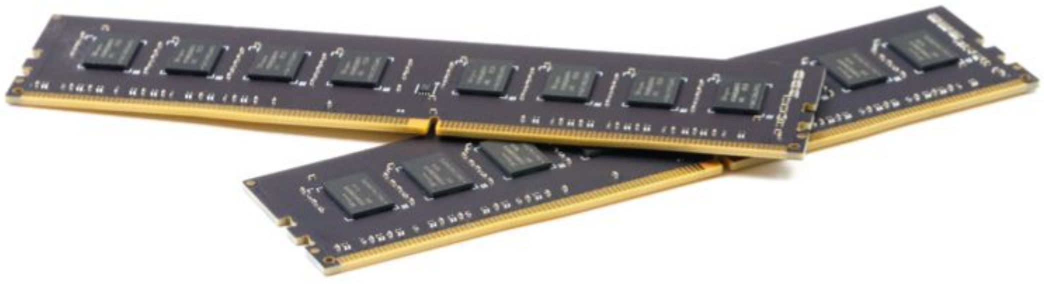
Training Pipeline: **Need Lots of Data**



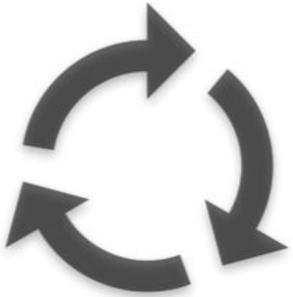
1000 Classes

1000 Images / Class

Training Pipeline: **Need Compute Resources**



Memory



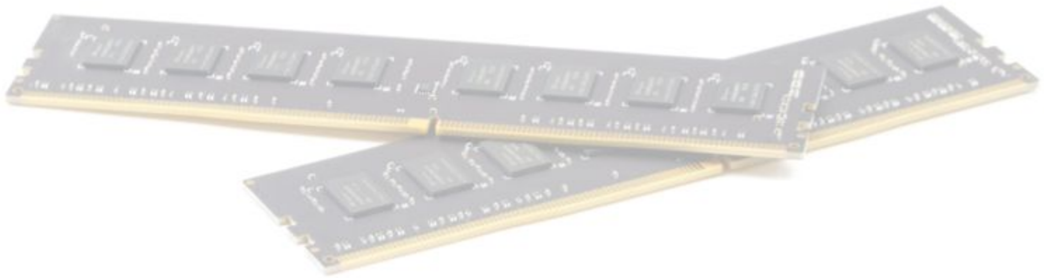
Compute

GPU and
Accelerators

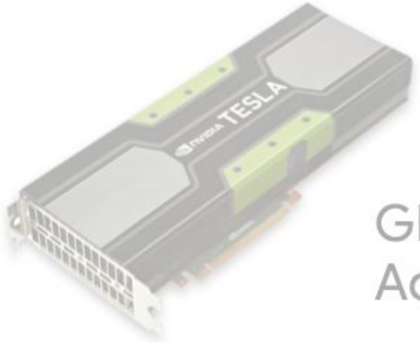
Training Pipeline: **Need Compute Resources**

Computationally Intensive
Repeated Many Times (Epochs)

Compute

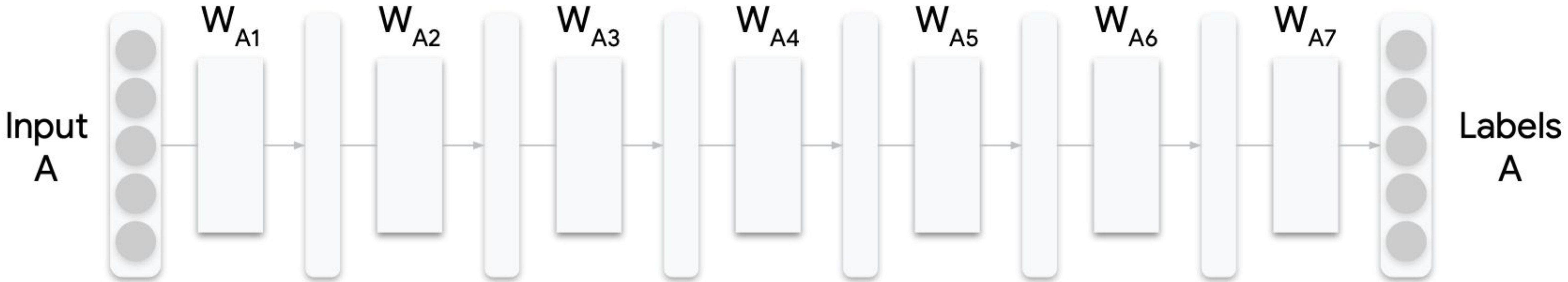


Memory

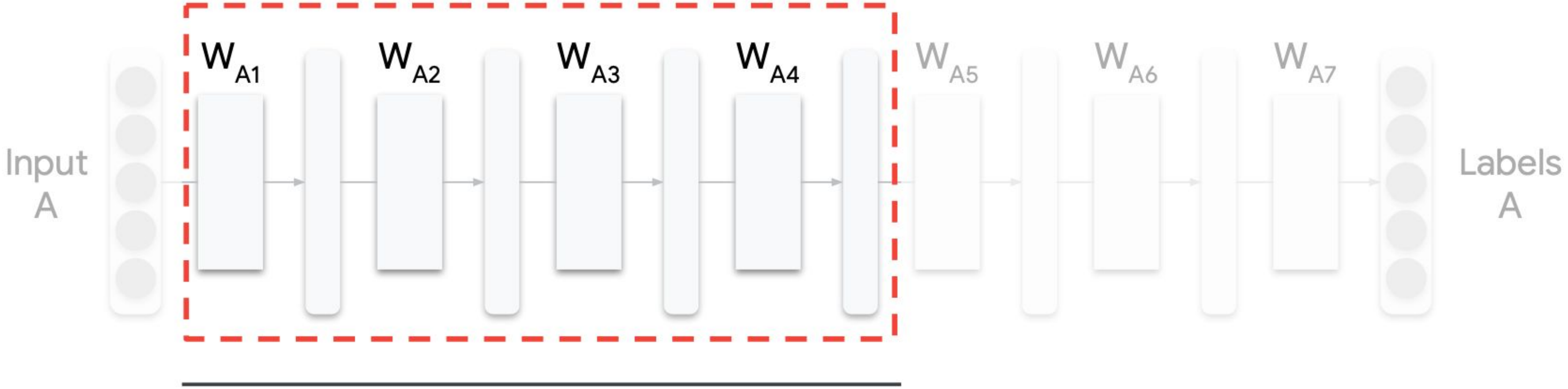


GPU and
Accelerators

End Result of Training

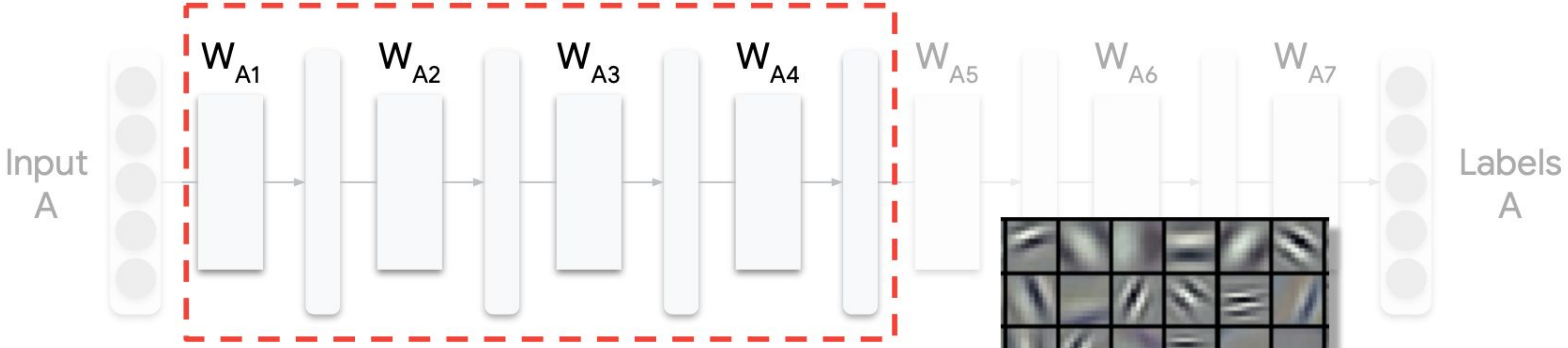


End Result of Training

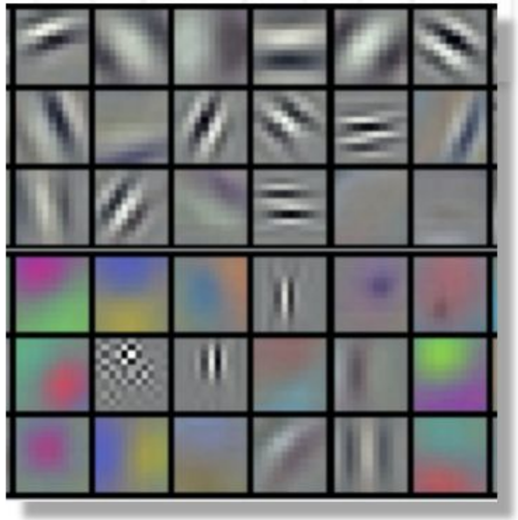


Learns ***general features***
irrespective of task

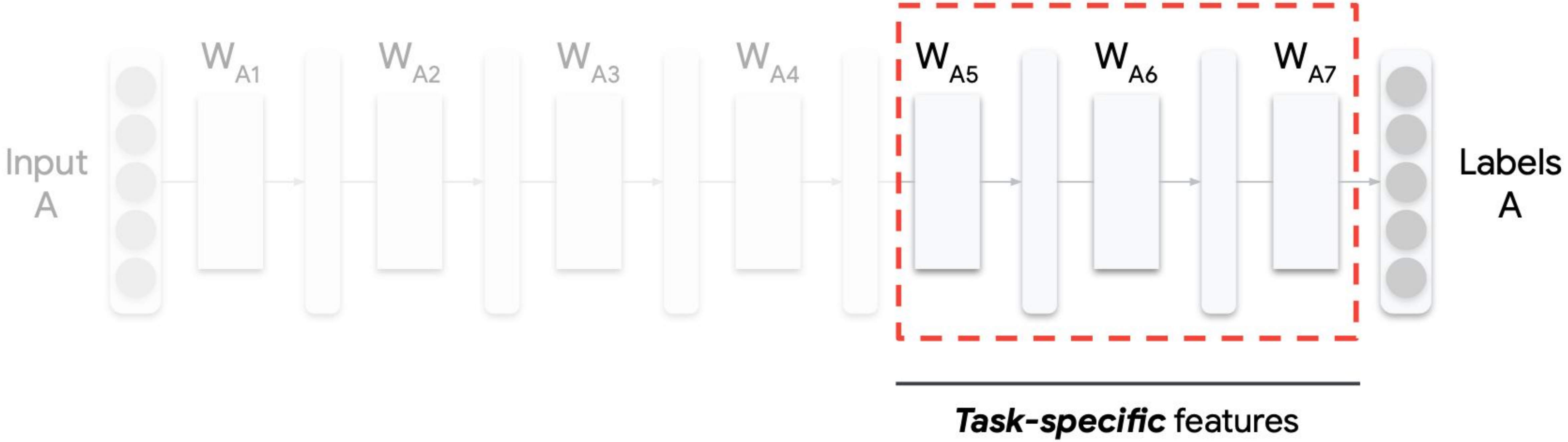
End Result of Training



Learns ***general features***
irrespective of task

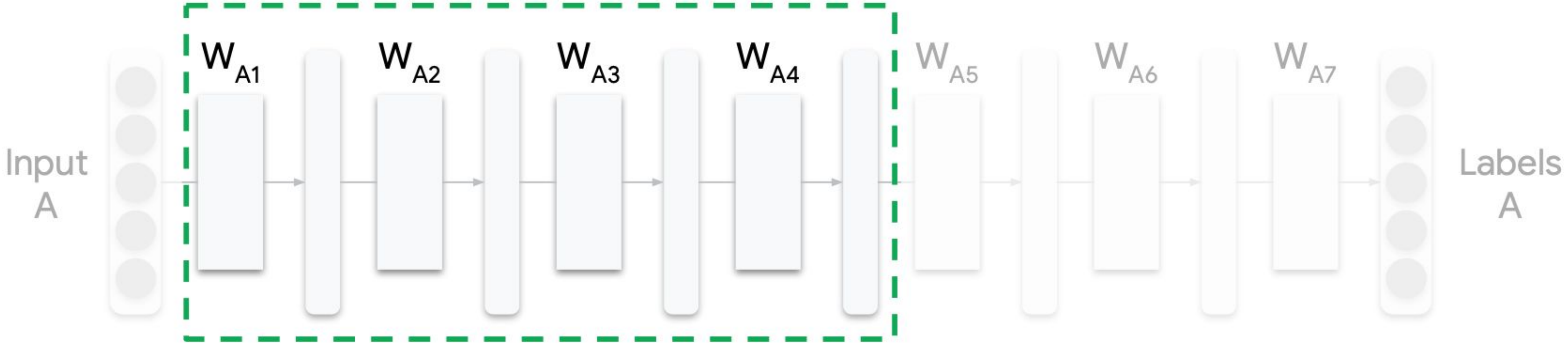


End Result of Training



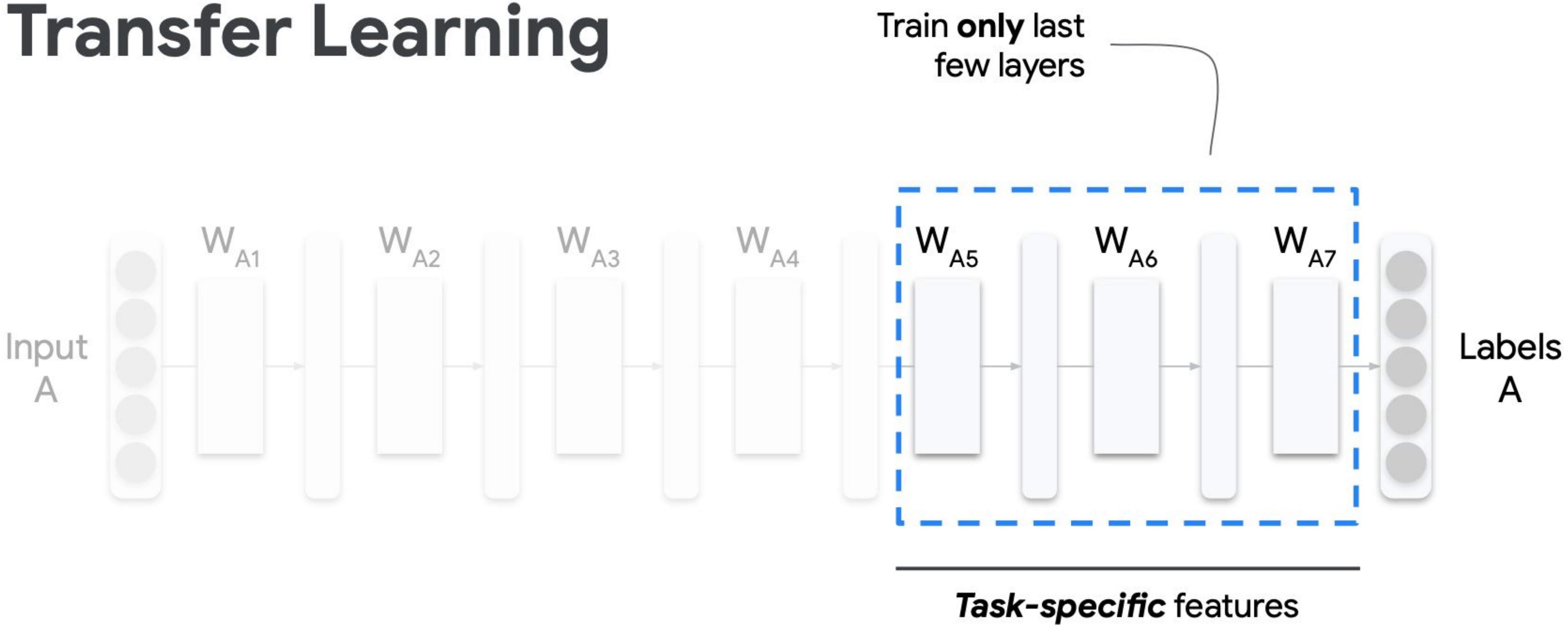
Transfer Learning

Reuse (freeze general feature extraction)



Learns ***general features*** irrespective of task

Transfer Learning

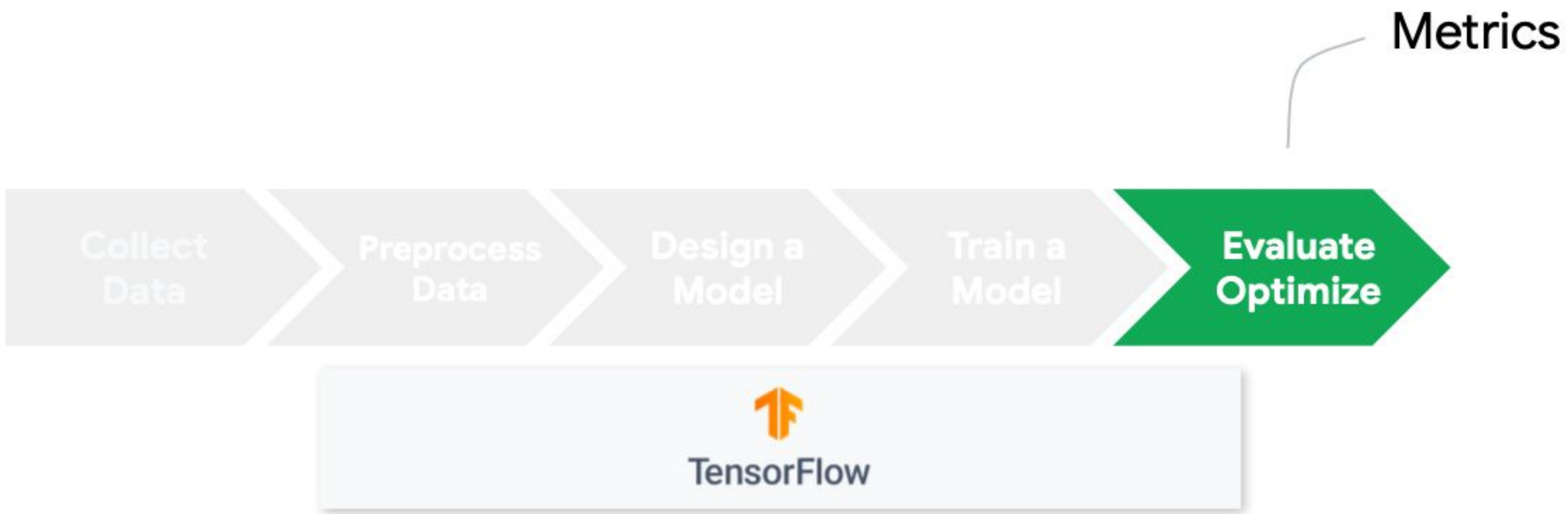


Mask Detection using Transfer Learning

Code Time!

Mask_Detection_using_Transfer_Learning.ipynb

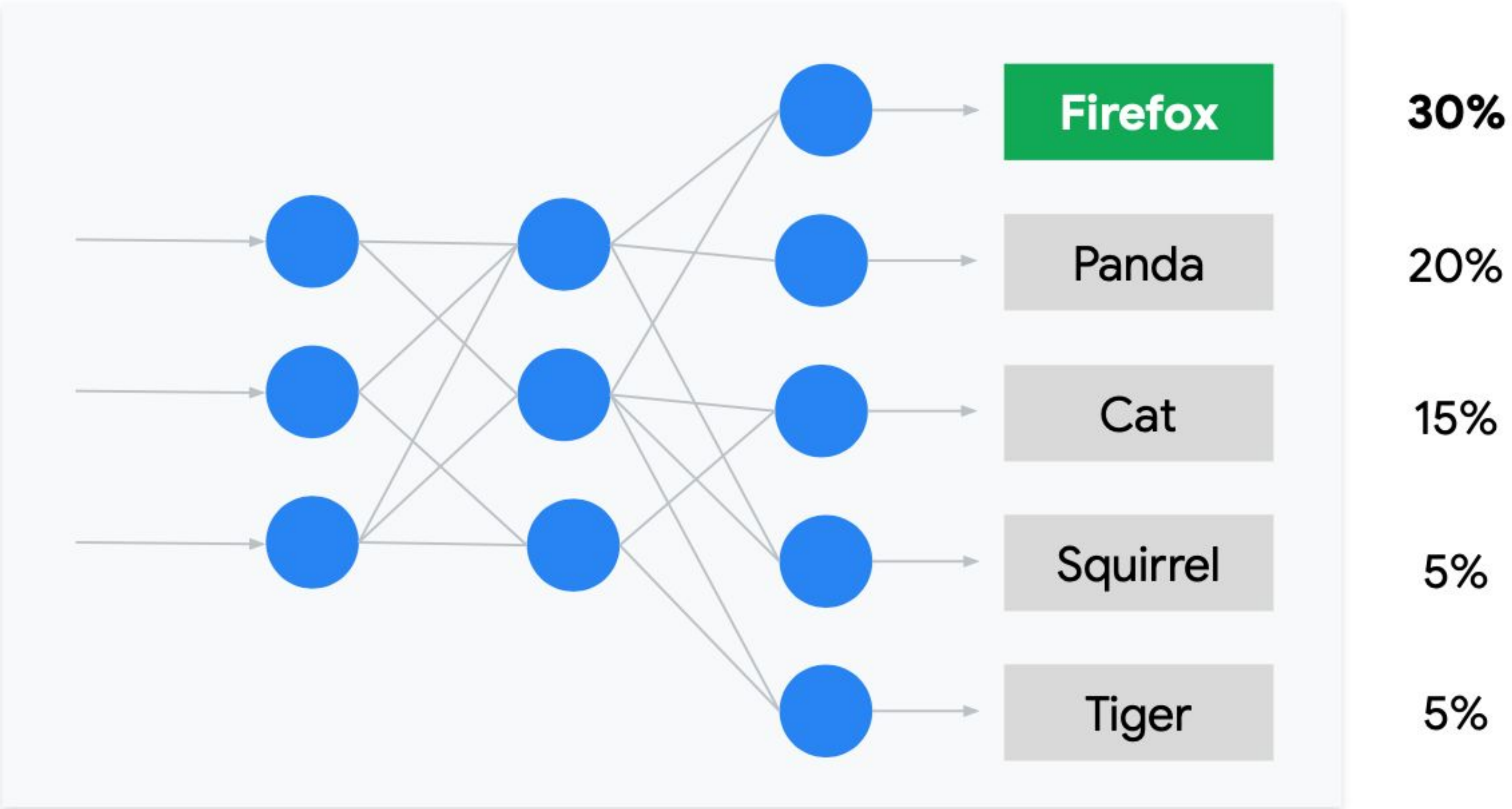
Visual Wake Words, Metrics


Metrics
Collect Data
Preprocess Data
Design a Model
Train a Model
Evaluate Optimize
TensorFlow

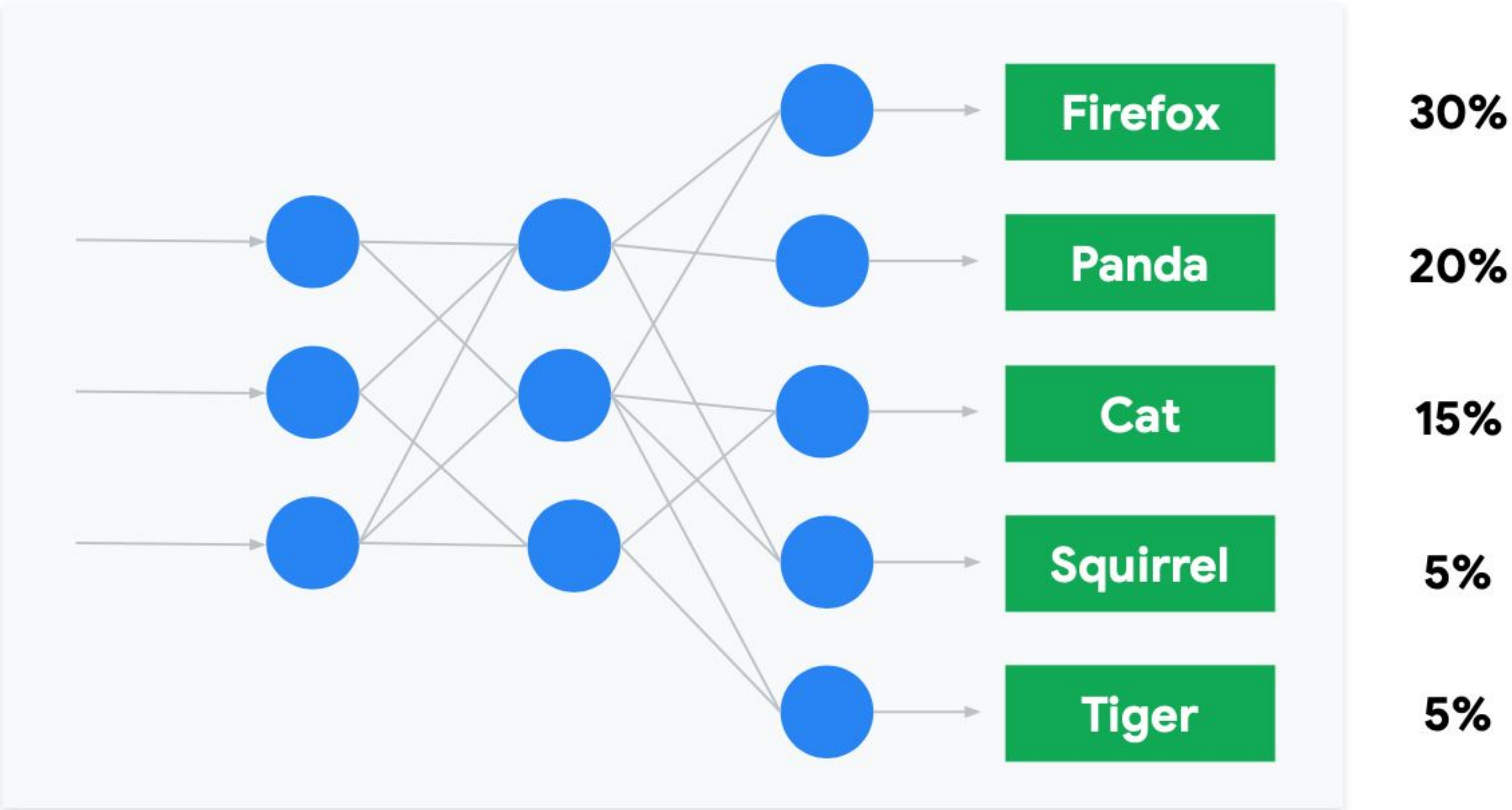
Common Metrics

Accuracy	Efficiency	User Experience
Quantitative	*Quantitative*	*Qualitative*

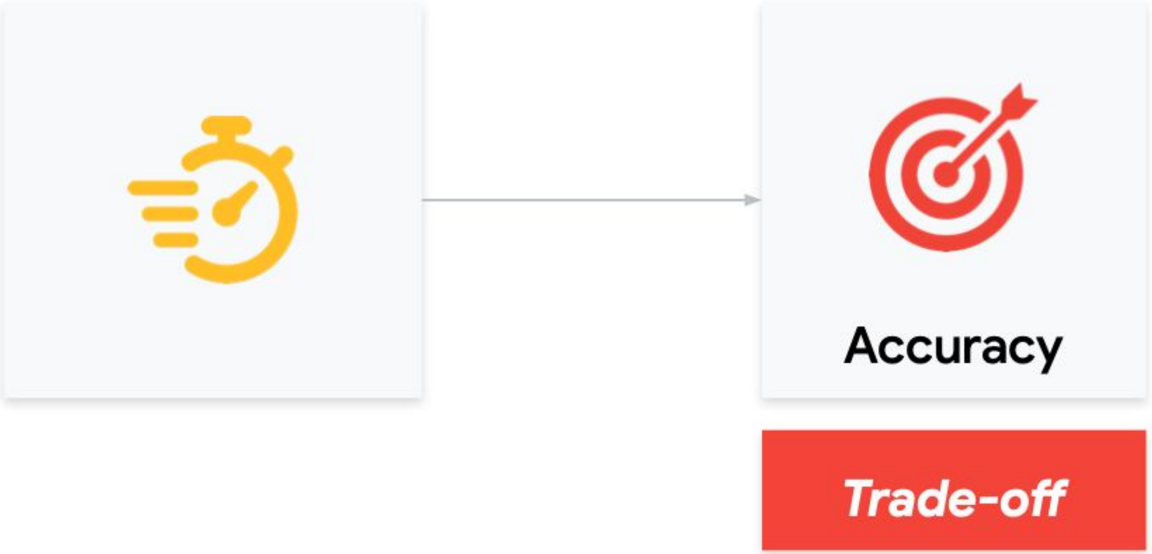
Top-1 Accuracy



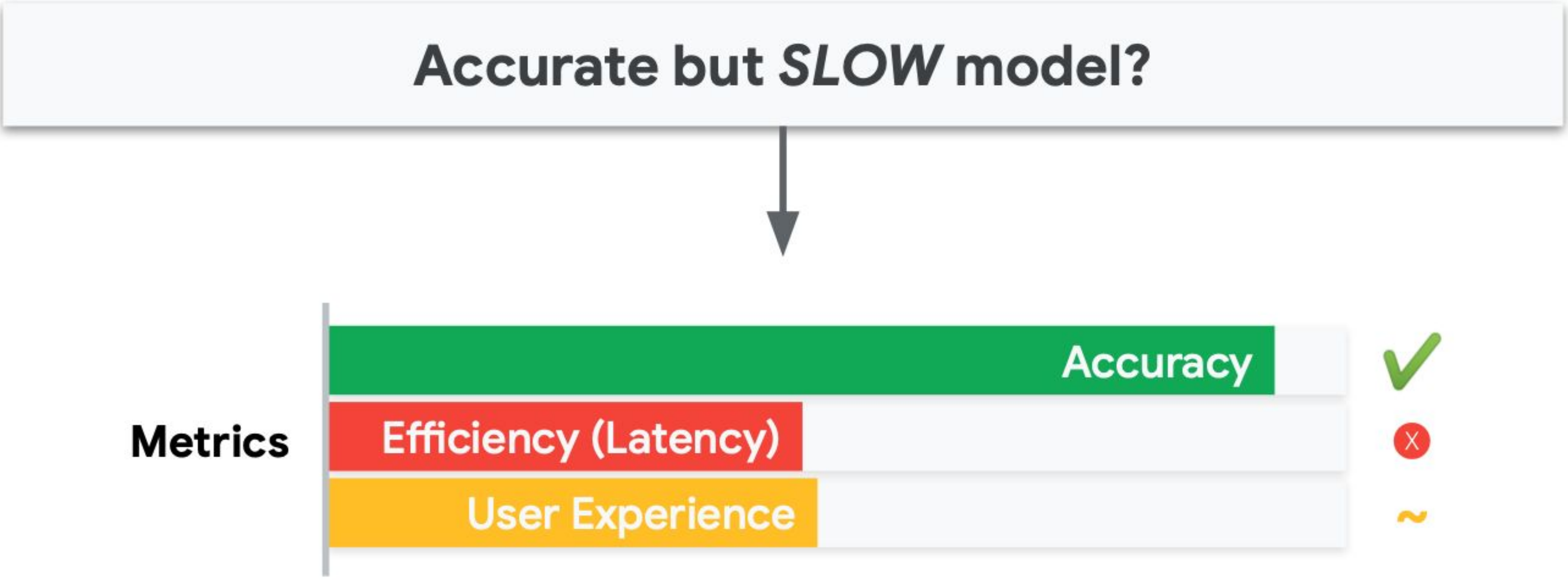
Top-5 Accuracy



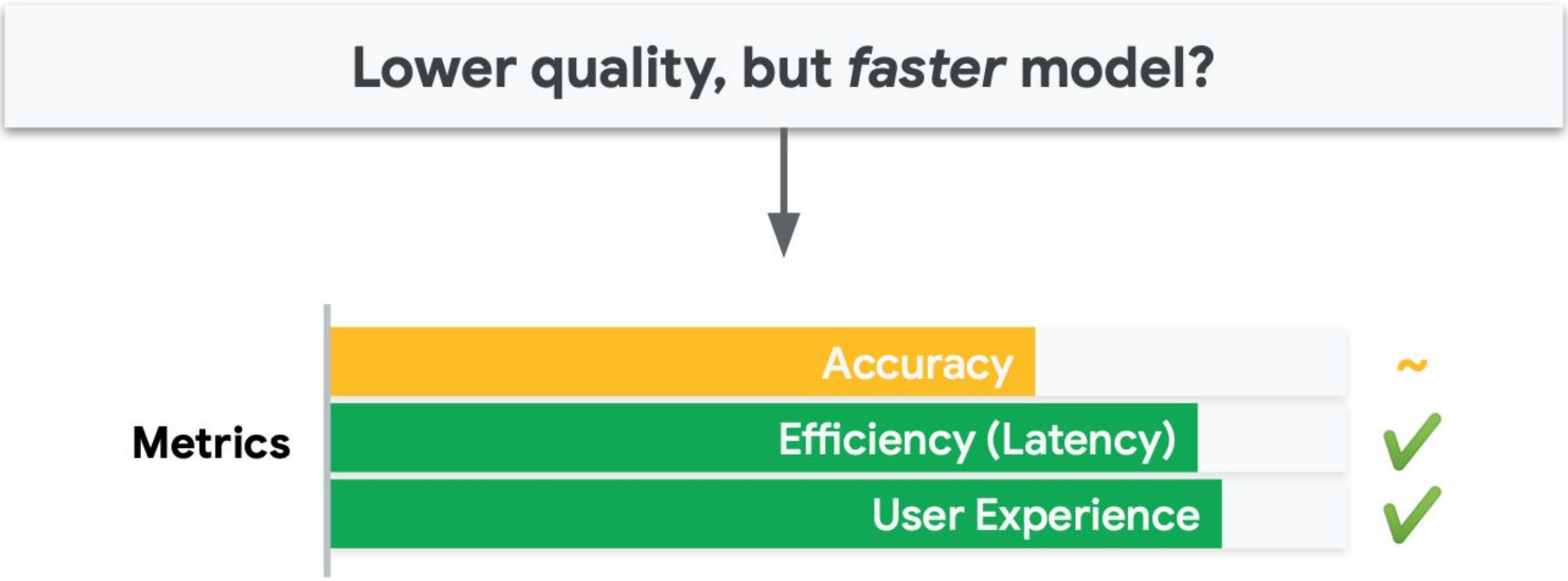
Latency



Latency


Accurate but *SLOW* model?
Metrics
Accuracy
Efficiency (Latency)
User Experience
X
~

Latency



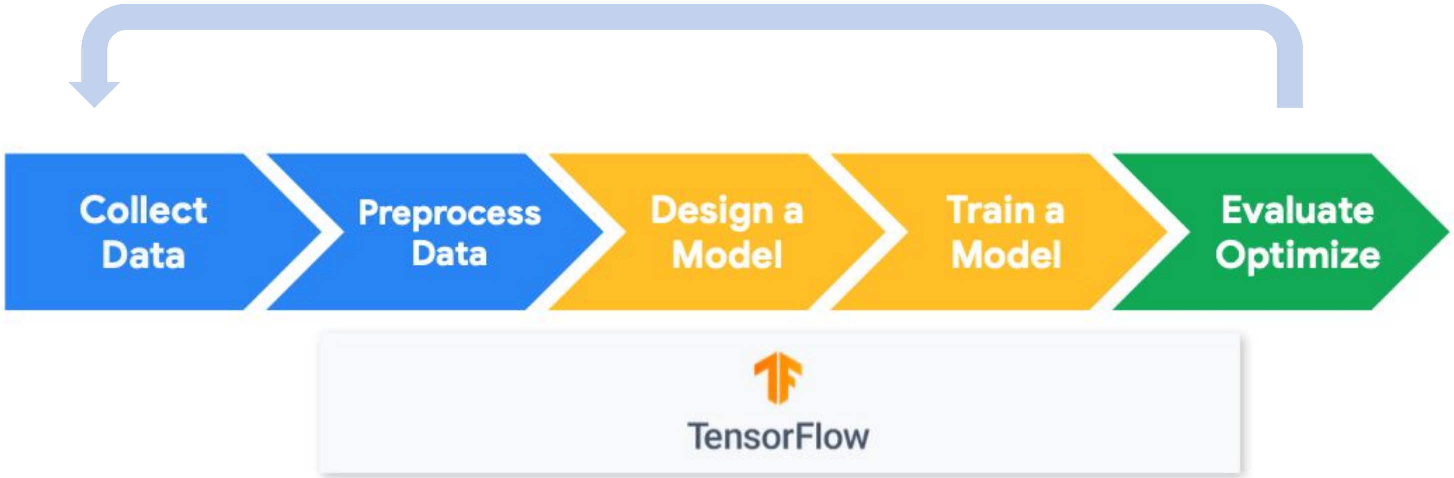
Lower quality, but *faster* model?
Metrics
Accuracy
~
Efficiency (Latency)
✓
User Experience
✓

Fairness

User in ***majority*** **group of training data?**

***Diverse*, representative data** is important because it enables fair use (equal performance) across populations

Achieving Ideal Metrics: Revisit pipeline



Reading Material

Main references

- Harvard School of Engineering and Applied Sciences - CS249r: Tiny Machine Learning
- Professional Certificate in Tiny Machine Learning (TinyML) – edX/Harvard
- Introduction to Embedded Machine Learning (Coursera)
- Text Book: "TinyML" by Pete Warden, Daniel Situnayake

I want to thank Shawn Hymel and Edge Impulse, Pete Warden and Laurence Moroney from Google, and especially Harvard professor Vijay Janapa Reddi, Ph.D. student Brian Plancher and their staff for preparing the excellent material on TinyML that is the basis of this course at UNIFEI.

The IESTI01 course is part of the TinyML4D, an initiative to make TinyML education available to everyone globally.

Thanks

And stay safe!

UNIFEI