Benchmarking Deep Learning

Rohith Bakkannagari
Senior Performance Engineer
The MathWorks

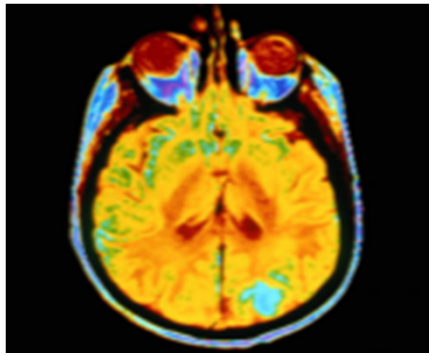
Agenda

- Deep Learning
- Challenges
- Deep Learning with MATLAB
- Deep Learning Workflows
- Benchmarking Results
- Summary

Deep Learning is Ubiquitous

Computer Vision

- Pedestrian and traffic sign detection
- Landmark identification
- Scene recognition
- Medical diagnosis and drug discovery



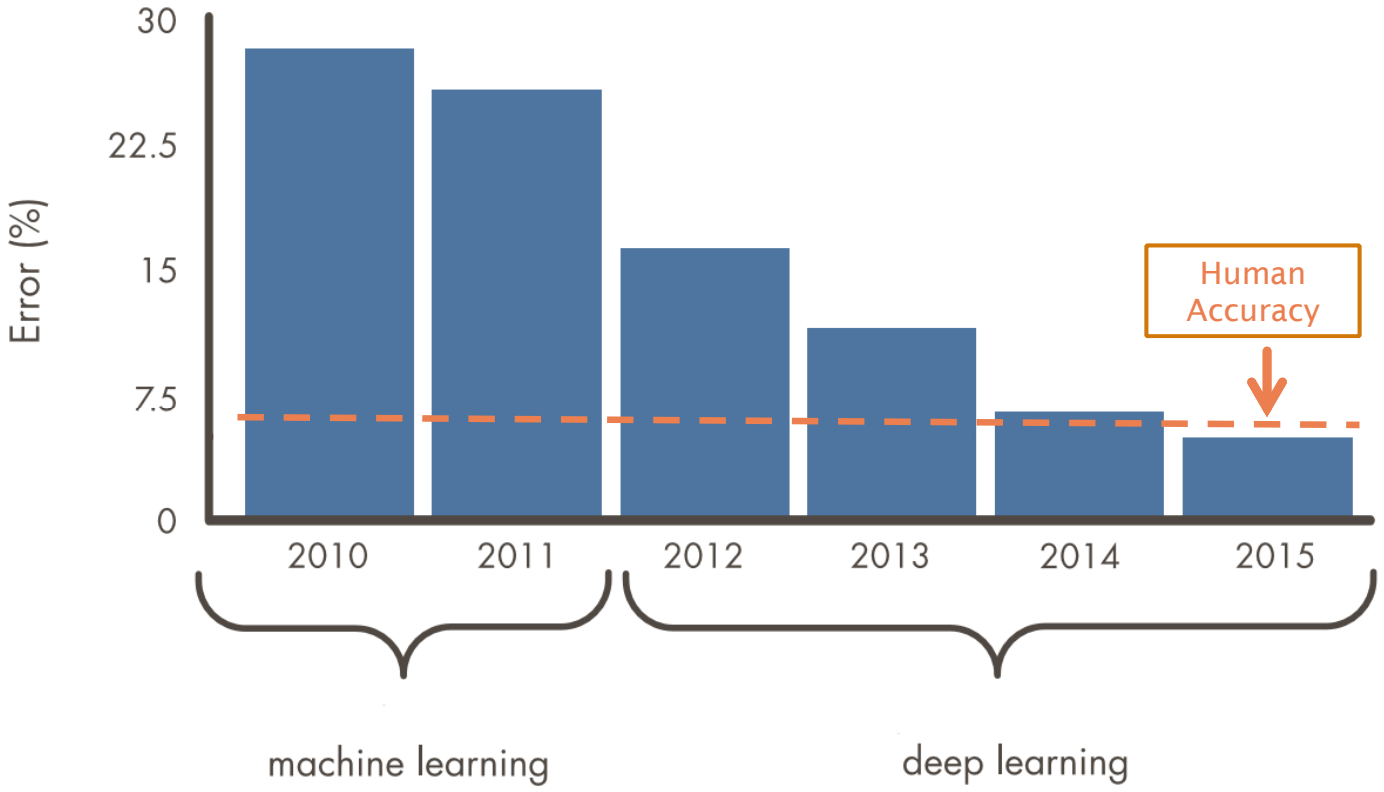
Text and Signal Processing

- Speech Recognition
- Speech & Text Translation

Robotics & Controls

and many more…

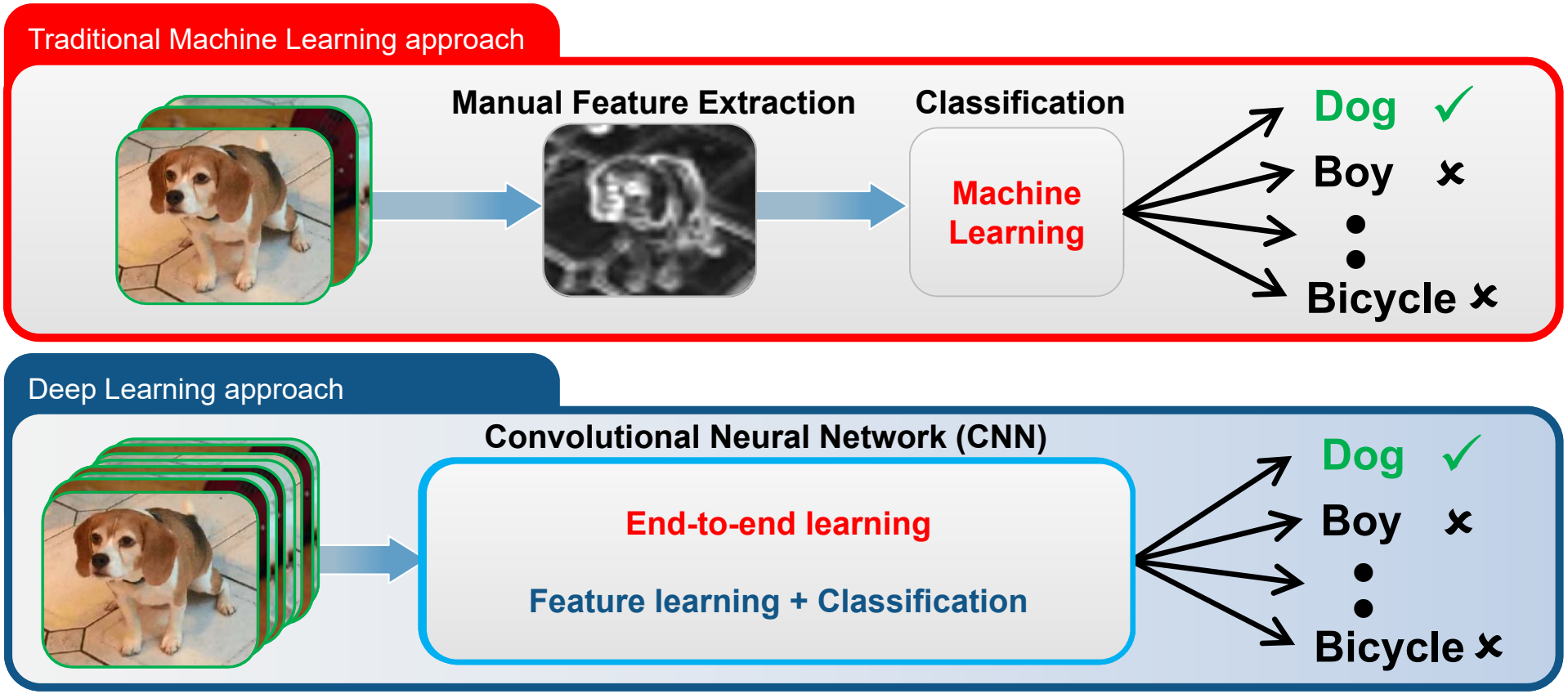
Why is Deep Learning So Popular Now?



Source: ILSVRC Top-5 Error on ImageNet

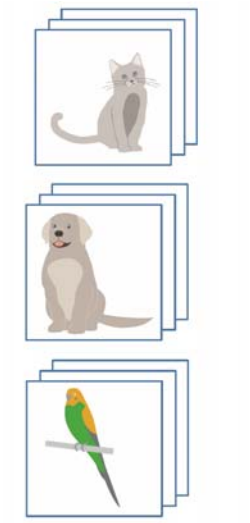
Deep Learning

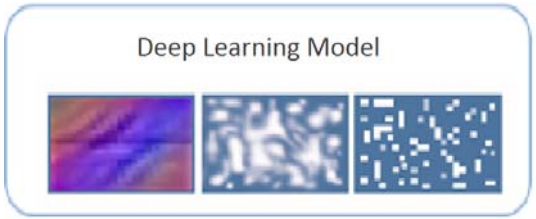
Deep learning is a **machine learning** technique that can learn **useful representations or features** directly from **images, text and sound**



What Is Deep Learning?

Deep learning learns **tasks** directly **from images, text, and sound.**

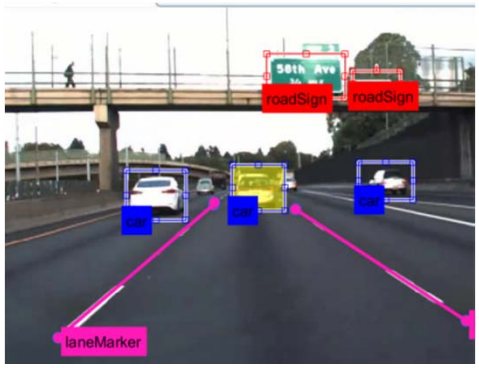




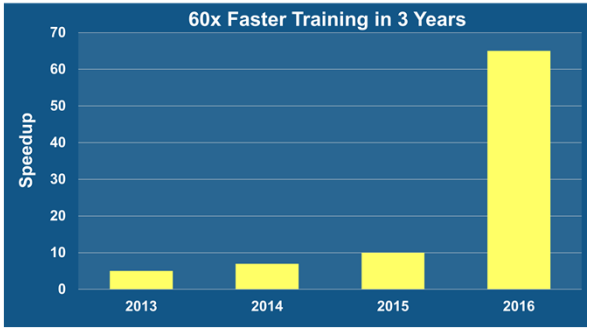
Data | **Models** | **Train / Run** | **Deploy / Share**

Why is Deep Learning Adoption Growing?

1. Massive Data Sets



2. Computing Power – High Performance GPUs



3. State-of-the-Art Models

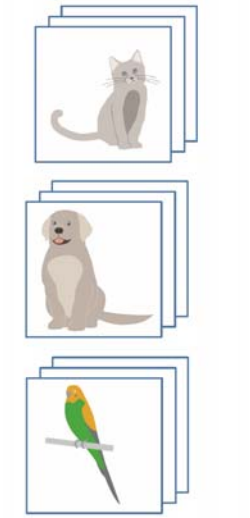
AlexNet
PRETRAINED MODEL

VGG-16
PRETRAINED MODEL

Caffe
MODELS

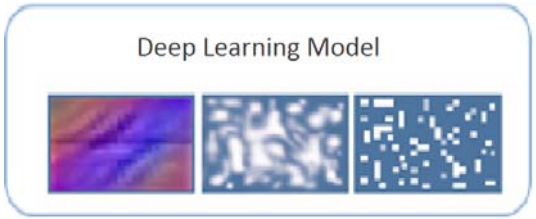
Deep Learning Challenges

"How do I *label* my data?"



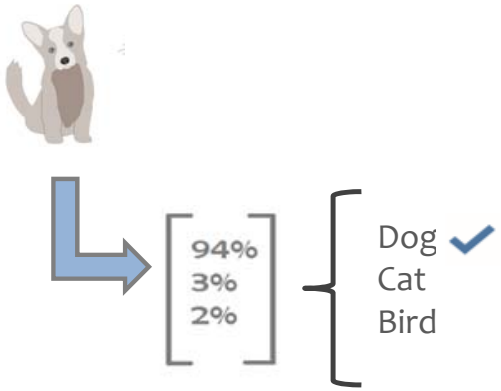
Data

"How do I *access* the latest models?"



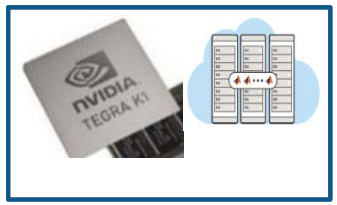
Models

"How do I make training and prediction *faster*?"



Train / Predict

"How do I *deploy* my new model?"



Deploy / Share

Deep Learning with MATLAB

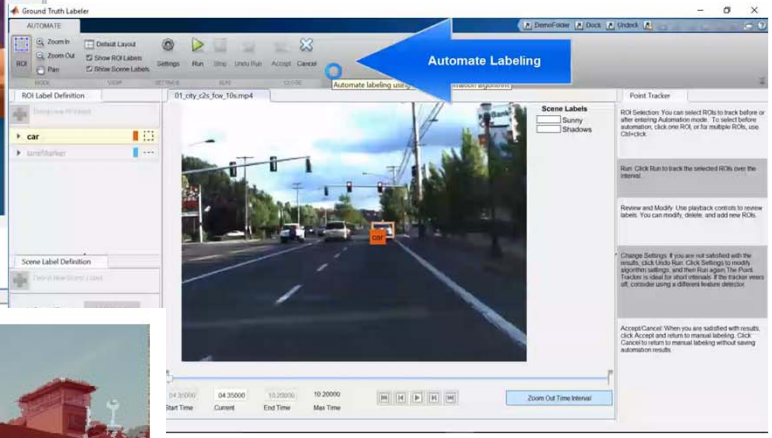
"How do I *label* my data?"	"How do I *access* the latest models?"	"How do I make training and prediction faster?"	"How do I deploy my new model?"
New App for Ground Truth Labeling Label pixels and regions for semantic segmentation	Caffe model importer LSTM (time series, text) DAG Networks Library of pretrained models	Multi-GPUs in parallel Optimized GPU code Training plots	**NEW PRODUCT:** **GPU Coder- Convert to NVIDIA CUDA code**
Data	**Models**	**Train / Predict**	**Deploy / Share**

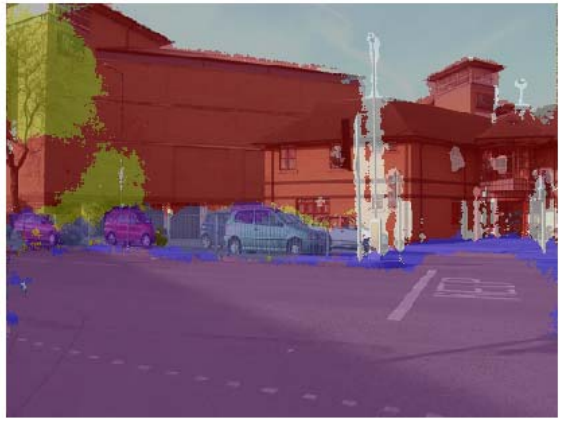
Deep Learning with MATLAB

"How do I *label* my data?"

New App for Ground Truth Labeling

Label pixels and regions for semantic segmentation





Data

Deep Learning with MATLAB

"How do I *access* the latest models?"

Caffe model importer

LSTM
(time series, text)

DAG Networks

Models

Pretrained Models
from Deep Learning Frameworks

AlexNet
PRETRAINED MODEL

VGG-16
PRETRAINED MODEL

VGG-19
PRETRAINED MODEL

Caffe
Caffe Model Zoo

Available Models:

- AlexNet
- VGG-16
- VGG-19
- GoogLeNet
- Many More with Caffe Model Importer

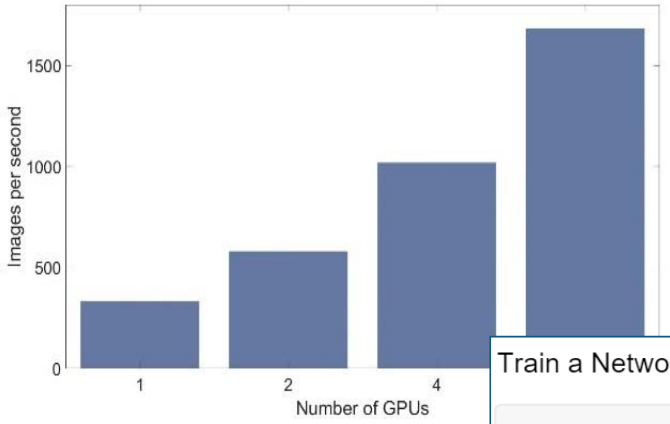
Deep Learning with MATLAB

"How do I make *training* faster?"

Multi GPUs in parallel

Training plots

Train

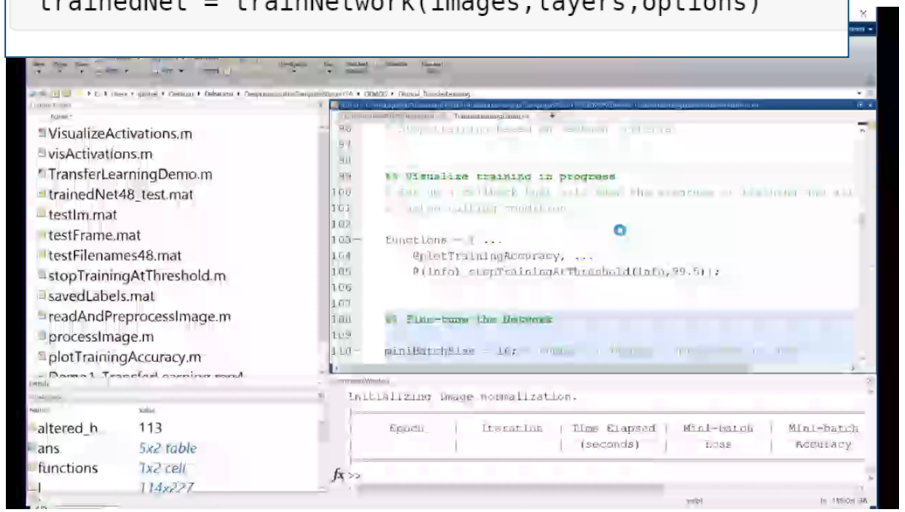


Training modes supported:

- Auto Select
- GPU
- Multi GPU (local)
- Multi GPU (cluster)

Train a Network:

```
trainedNet = trainNetwork(images,layers,options)
```

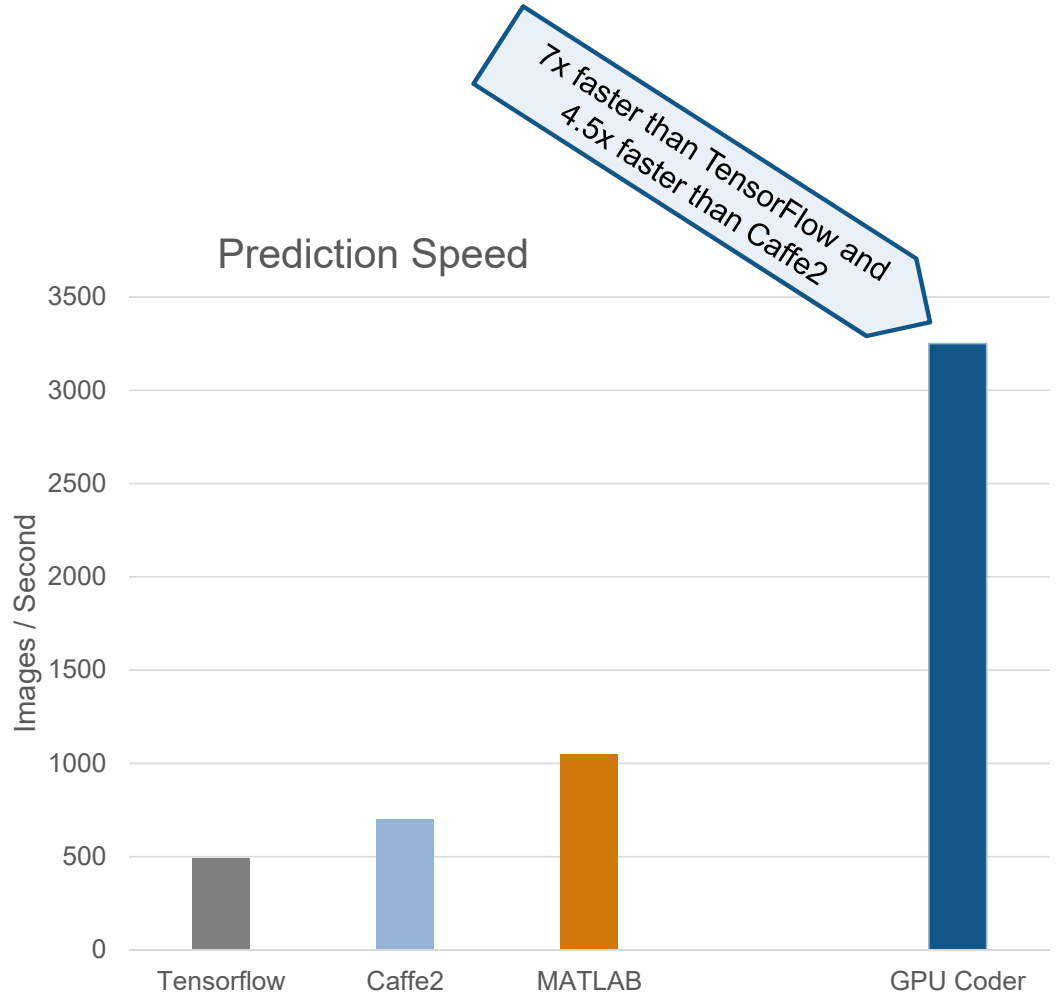


Running a Trained Model

R2017b

Introducing:

GPU Coder- Convert to NVIDIA CUDA code



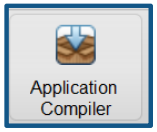
Deployment

"How do I *deploy* my model?"

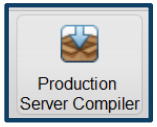
Introducing:

GPU Coder- Convert to NVIDIA CUDA code

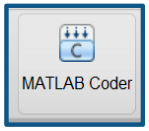
Deploy / Share



- Create Desktop Apps



- Run Enterprise Solution



- Generate C and C++ Code

- Target GPUs

GPU Coder

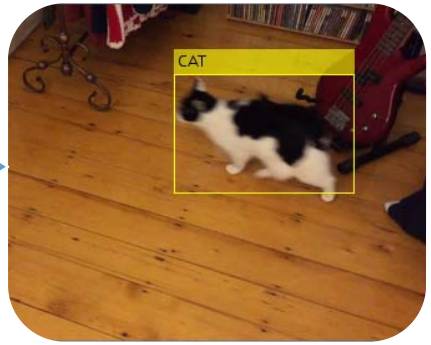
NVIDIA CUDA

Example: Pet Detection and Classification

Input:
Video containing
a cat or dog

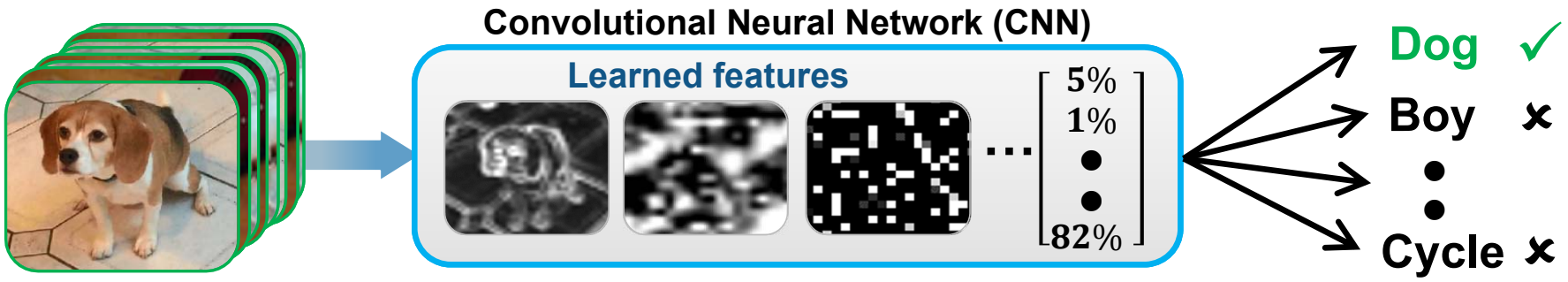
1. **Detection**
2. **Classification**

Output:
Detect and
recognize pet



Two Deep Learning Workflows

Train CNN from scratch



Recommended <u>only</u> when:

Training data	1000s to millions of labeled images
Computation	Compute intensive (requires GPU)
Training Time	Days to Weeks for real problems
Model accuracy	High (can over fit to small datasets)

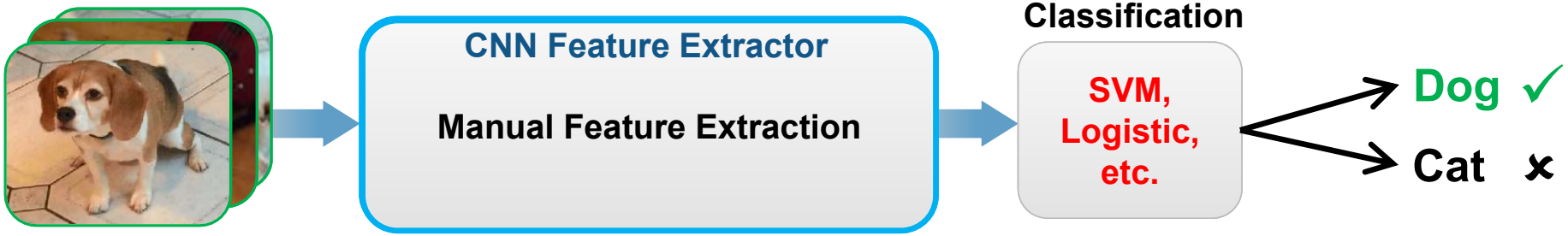
Two Deep Learning Workflows

Prediction or Inference

Use a pretrained CNN as an automatic feature extractor

Recommended when:

Training data	100s to 1000s of labeled images (small)
Computation	Moderate computation (GPU optional)
Training Time	Seconds to minutes
Model accuracy	Good, depends on the pretrained CNN model



We ran benchmarks on a workstation with latest Titan Xp GPU

Hardware

NVIDIA Titan Xp GPU used – highest end GPU available

Prediction

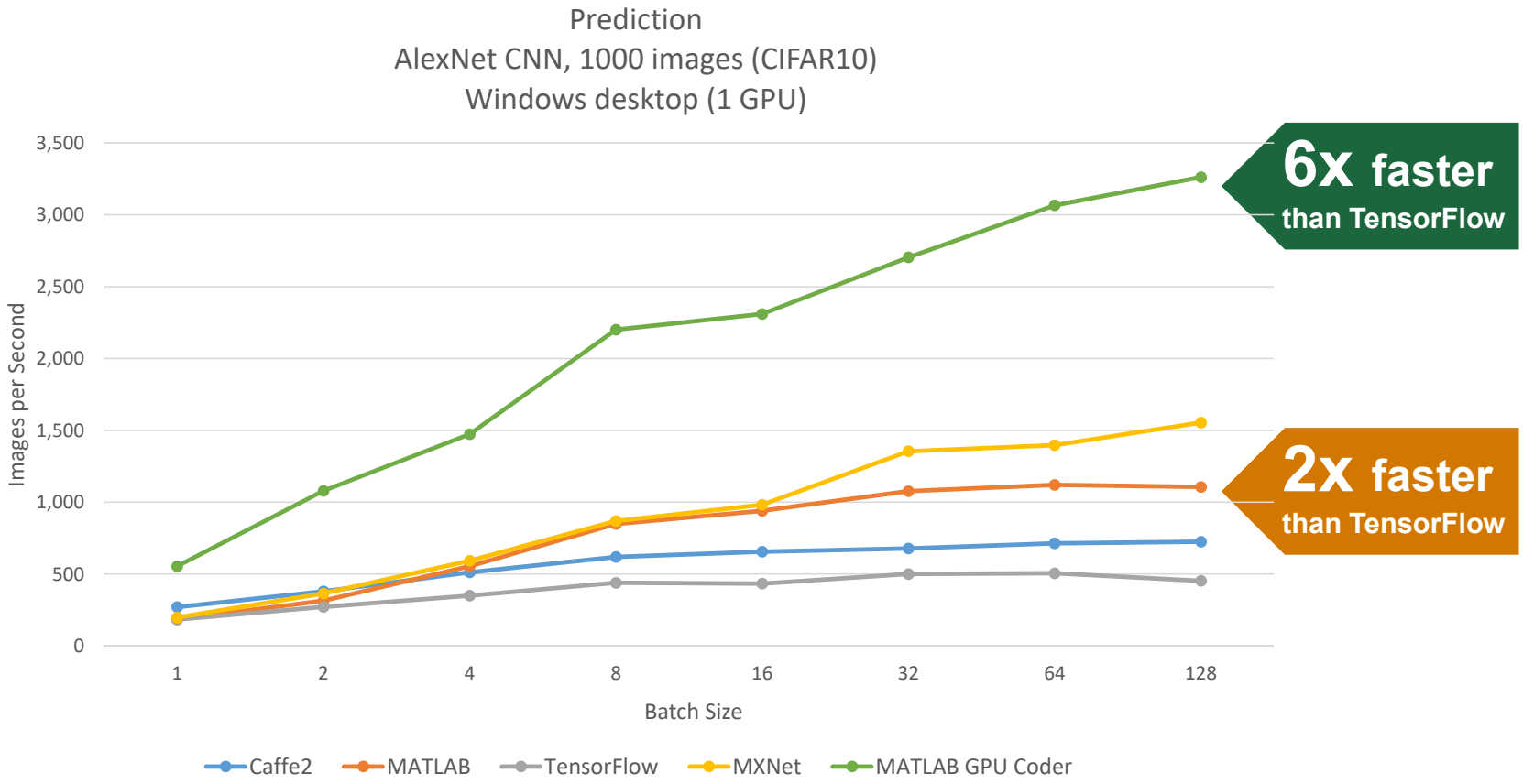
- Recorded images/sec with varied batch size
- GPU and CPU data

Training

- Recorded time per epoch and final accuracy score
- Trained for 10-25 epochs

* *Internal benchmarks were performed for inference performance of AlexNet using a TitanXP GPU and Intel(R) Xeon(R) CPU E5-1650 v4 @ 3.60GHz. Software versions used were MATLAB(R2017b), TensorFlow(1.2.0), and Caffe2(0.8.1). The GPU accelerated versions of each software were used for benchmarks. All tests were run on Windows 10.*

Prediction (using GPU): MATLAB GPU Coder is the fastest, and MATLAB is twice as fast as TensorFlow



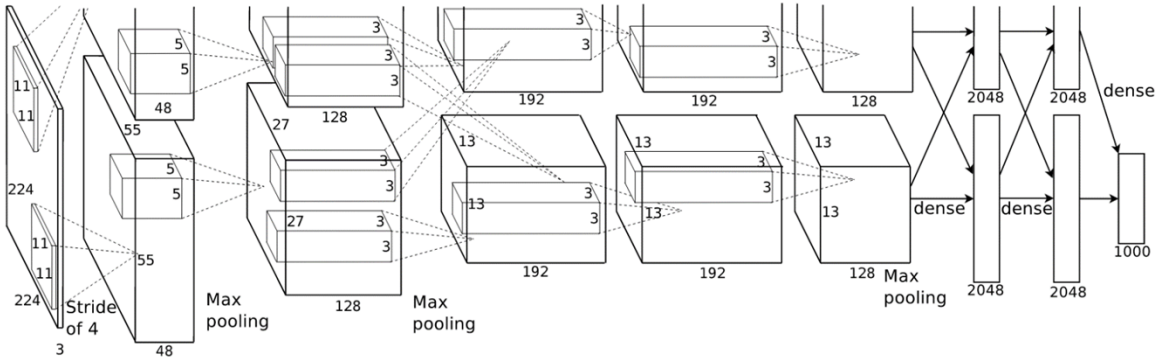
Our goal is to use AlexNet architecture for training from scratch

Popular AlexNet architecture from "**ImageNet Classification with Deep Convolutional Neural Networks**" by Krizhevsky et al.

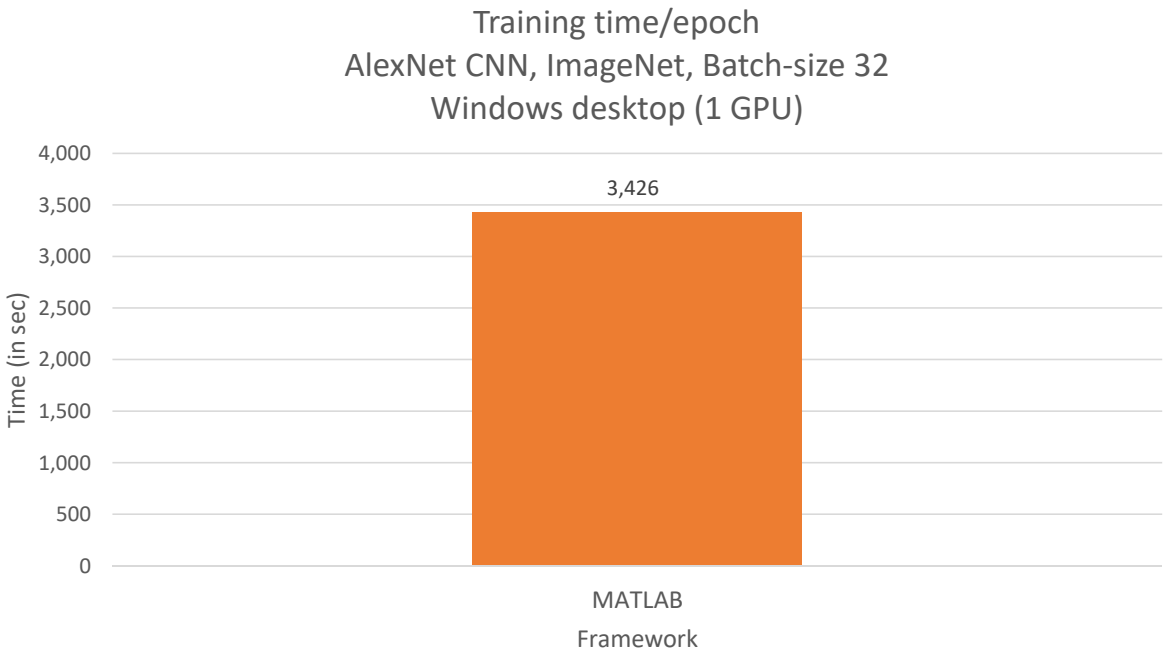
Used the ILSVRC2012 dataset (138GB)

- Subset of the popular ImageNet dataset (1.3TB)

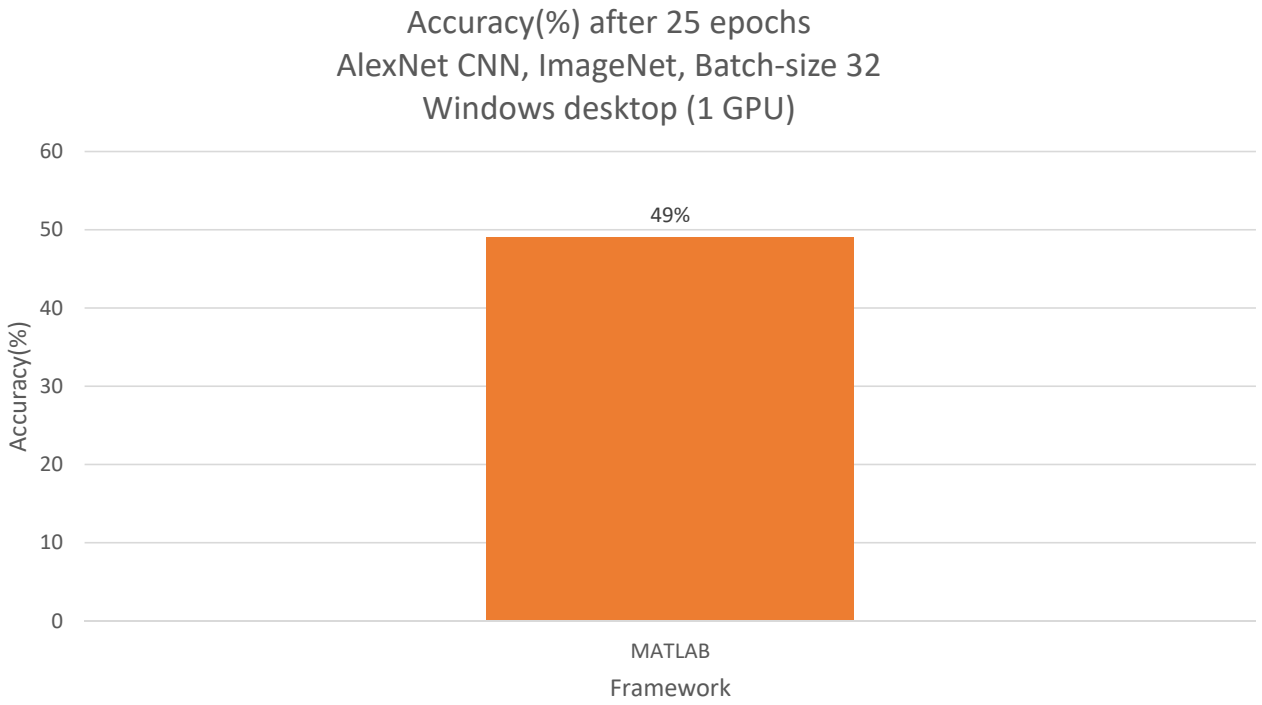
Had to standardize implementations across different frameworks



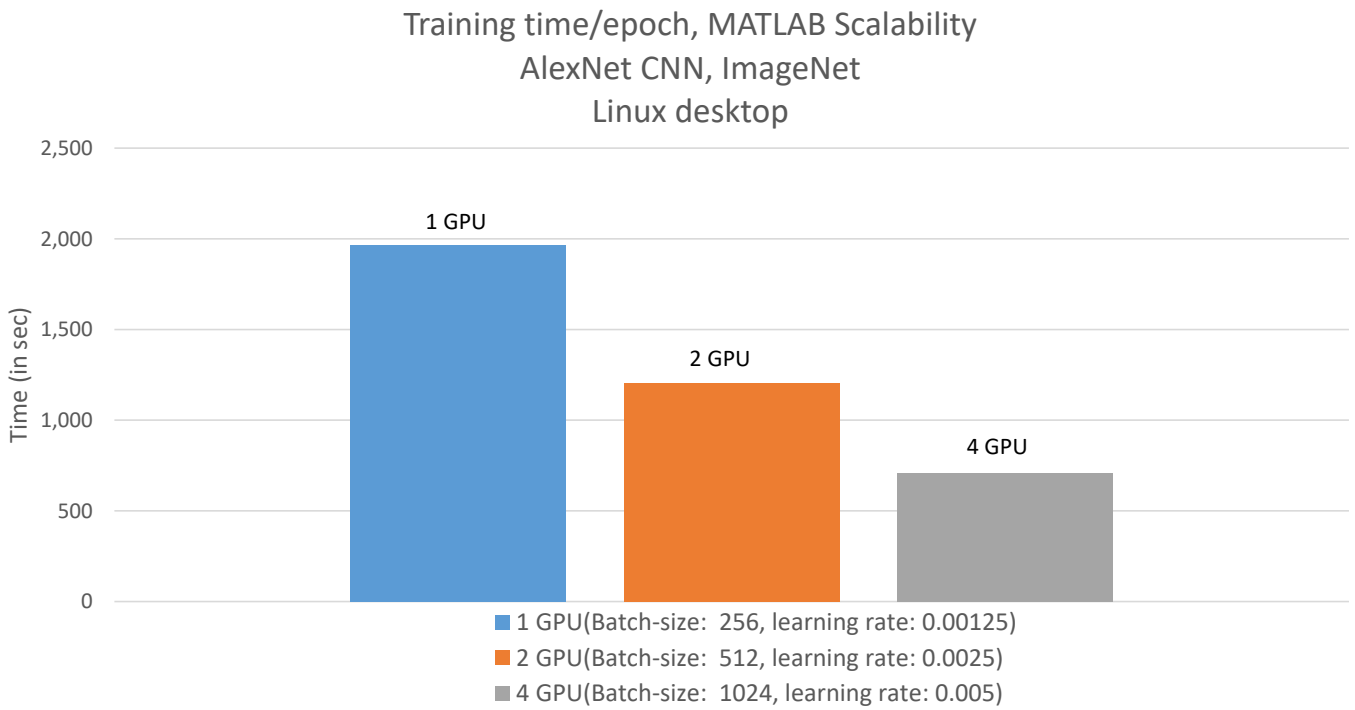
Training from Scratch: Performance



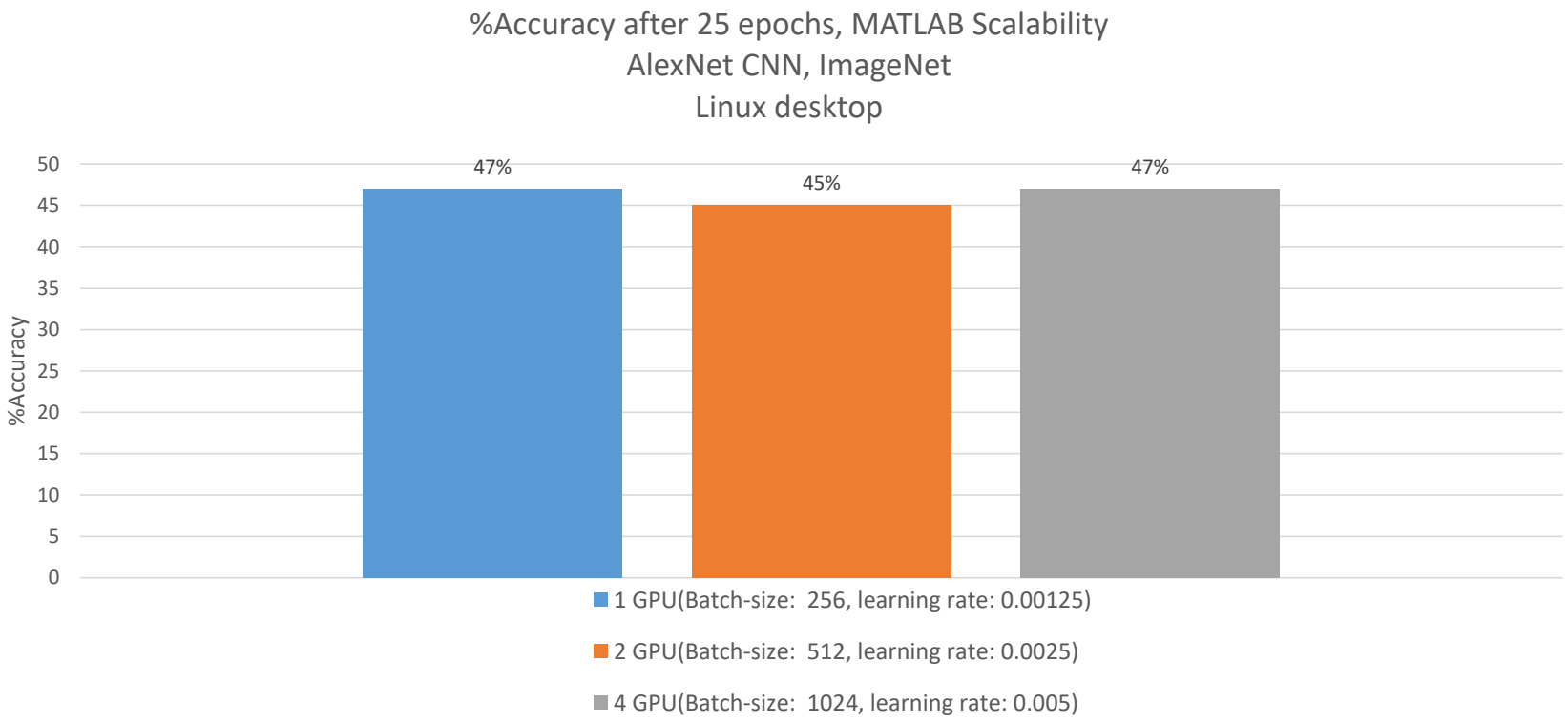
Training from Scratch: Accuracy



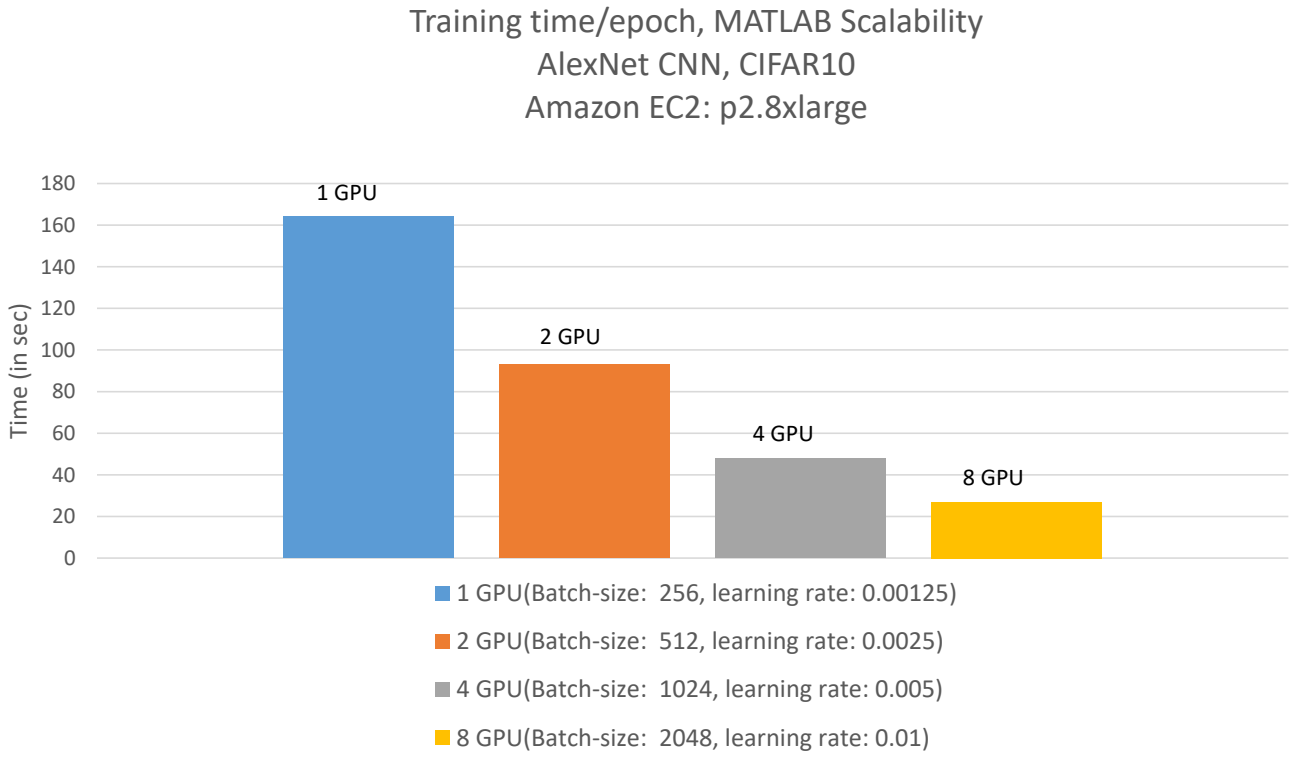
Training from Scratch: Training gets faster with more GPUs



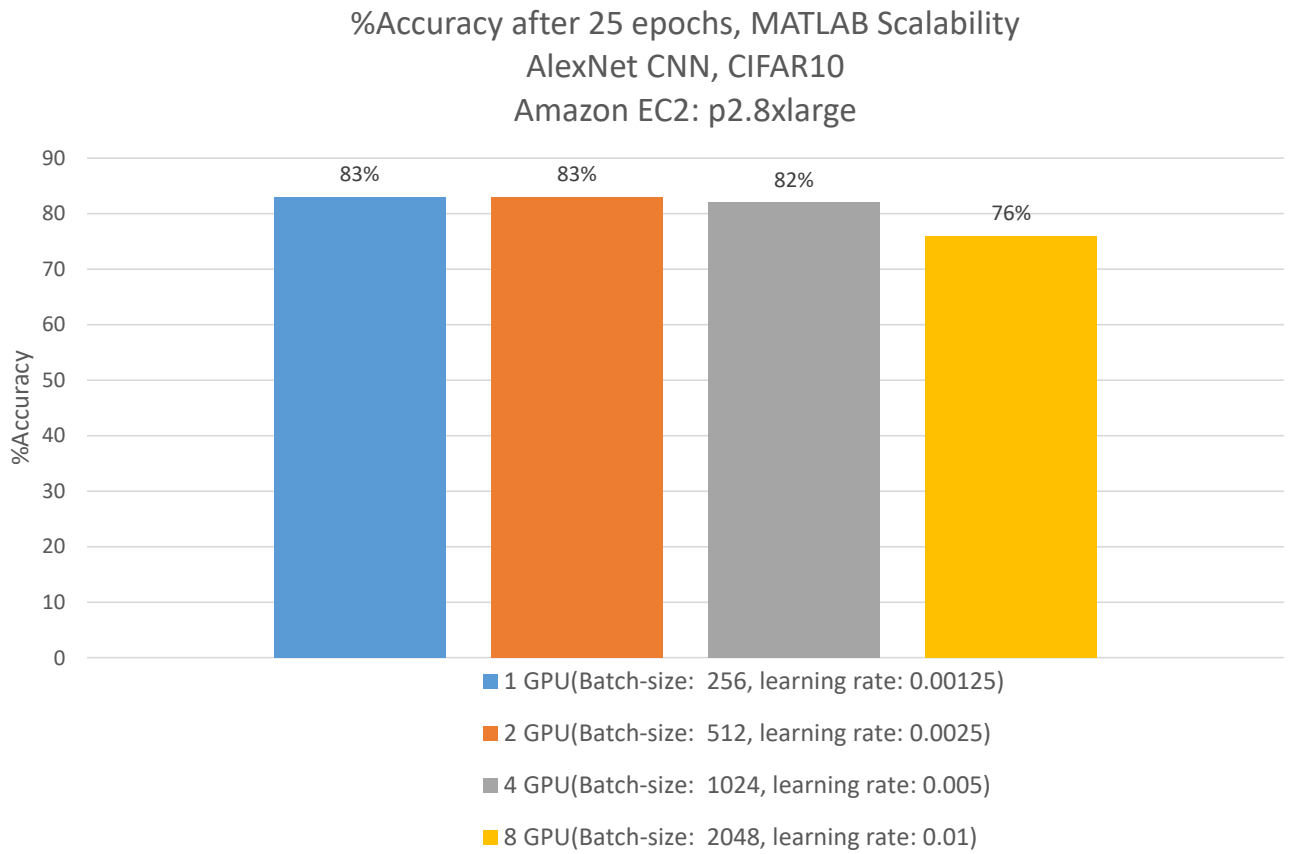
Training from Scratch: Accuracy of the trained model



Training from Scratch: Training gets faster with more GPUs (Amazon EC2: p2.8xlarge cluster)



Training from Scratch: Accuracy of the trained model (Amazon EC2: p2.8xlarge cluster)



Summary

- MATLAB makes it easy to design, build, train, and deploy models
- For prediction performance, MATLAB GPU Coder outperforms other deep learning frameworks
- For training performance, MATLAB scales well with multiple GPUs