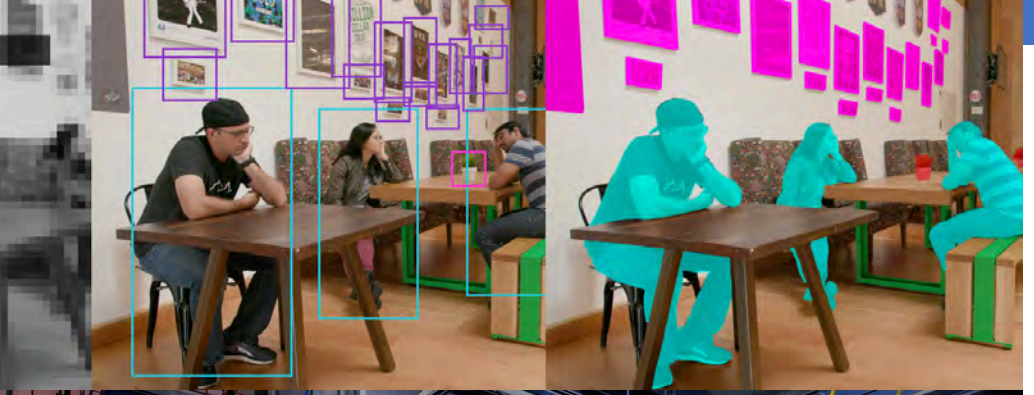


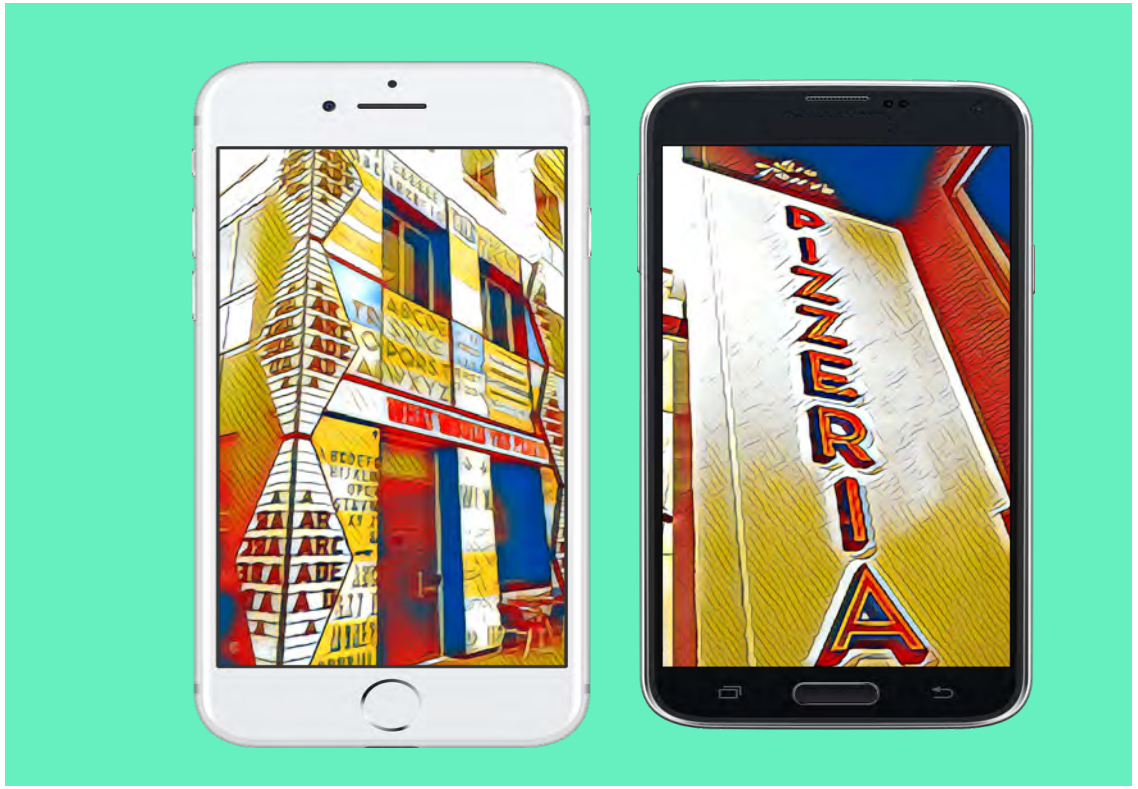
Caffe2

Caffe2 is …

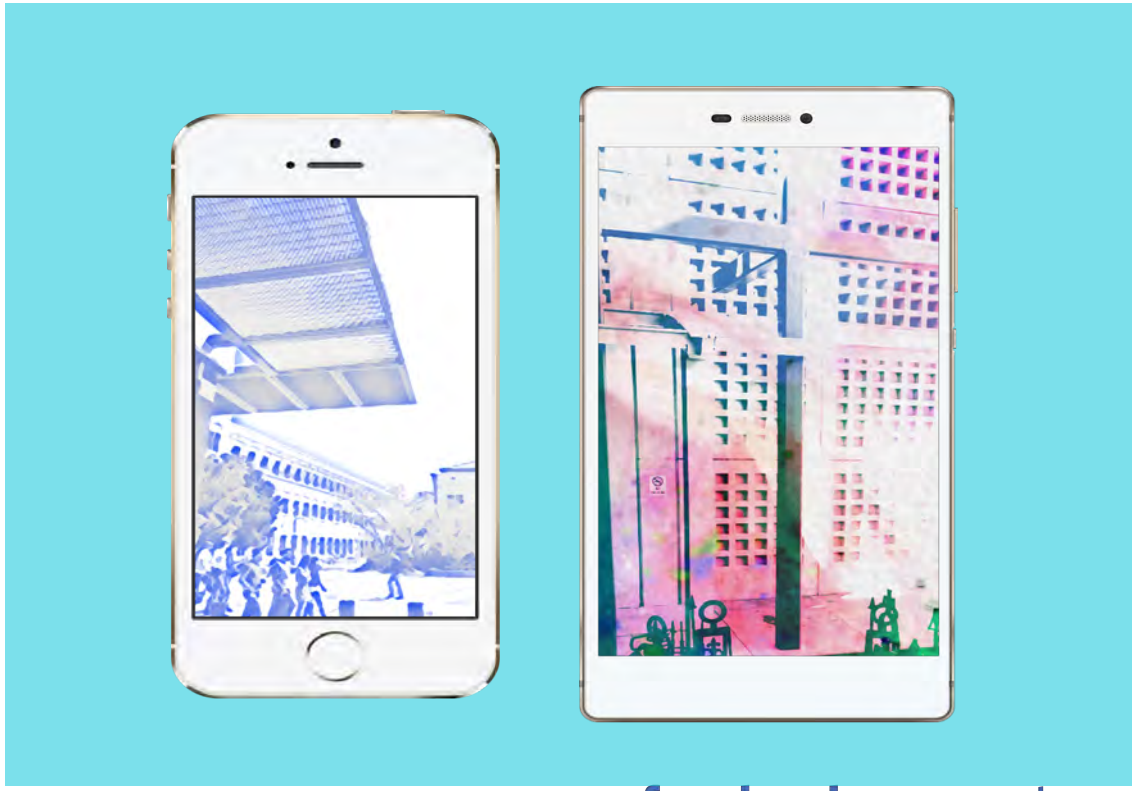
- A lightweight framework for deep learning algorithms
- Primarily designed for production use cases
  - Speed is top
  - C++ / Python based interfaces
- First-class distributed support
- Cross-platform
- CV / AR / NLP / Speech / Ranking apps

Optimized for Mobile Inference

Year Class 2015+
Optimized



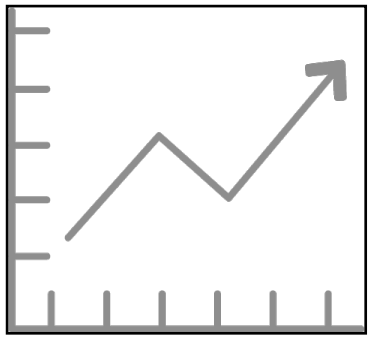
Year Class 2013, 2014
Supported



facebook research

Training ImageNet using Caffe2 (ResNet-50)

Bring down time to train a new image classifier from multiple days to ***a single hour***



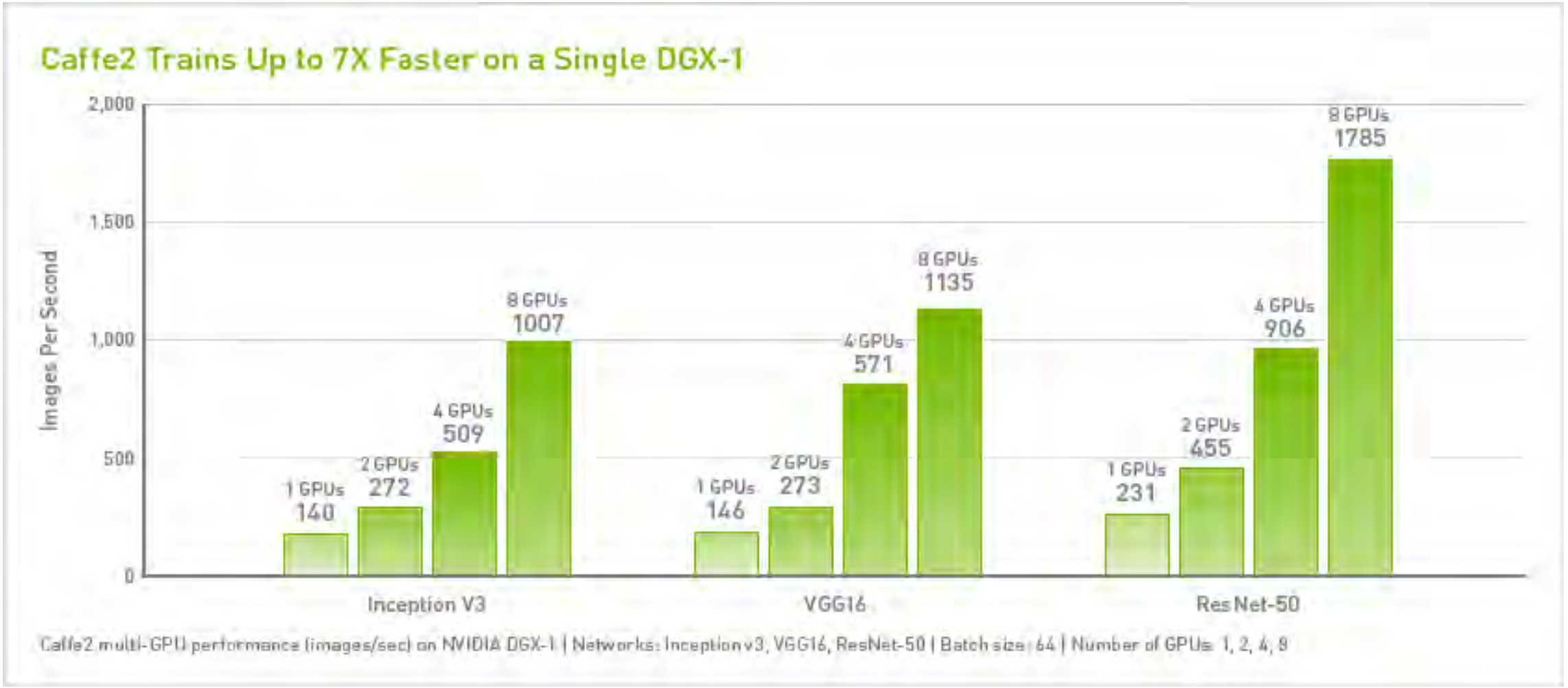
Scale up input batch size to 8192

Use 256 NVIDIA P100 GPUs

(90% scaling efficiency)

facebook research

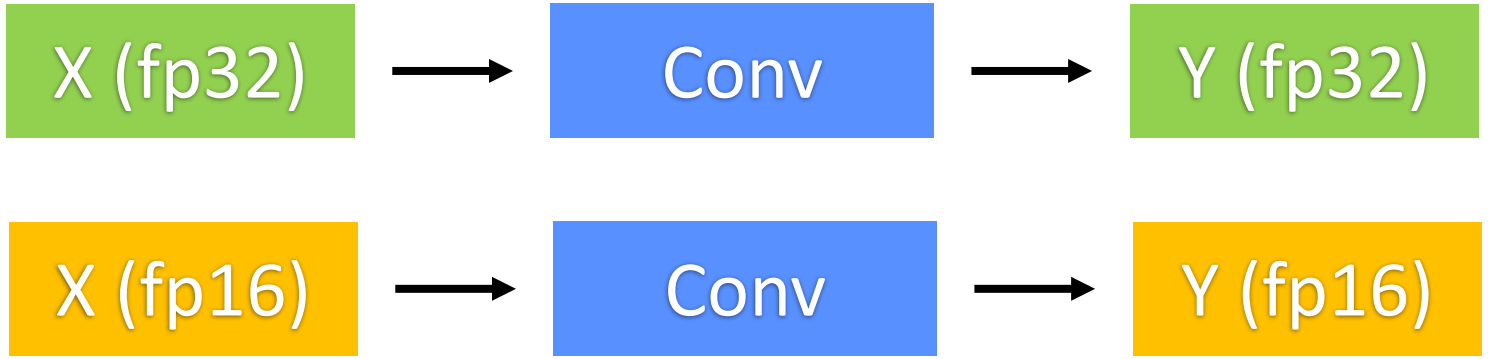
Scalable Distributed Training



facebook research

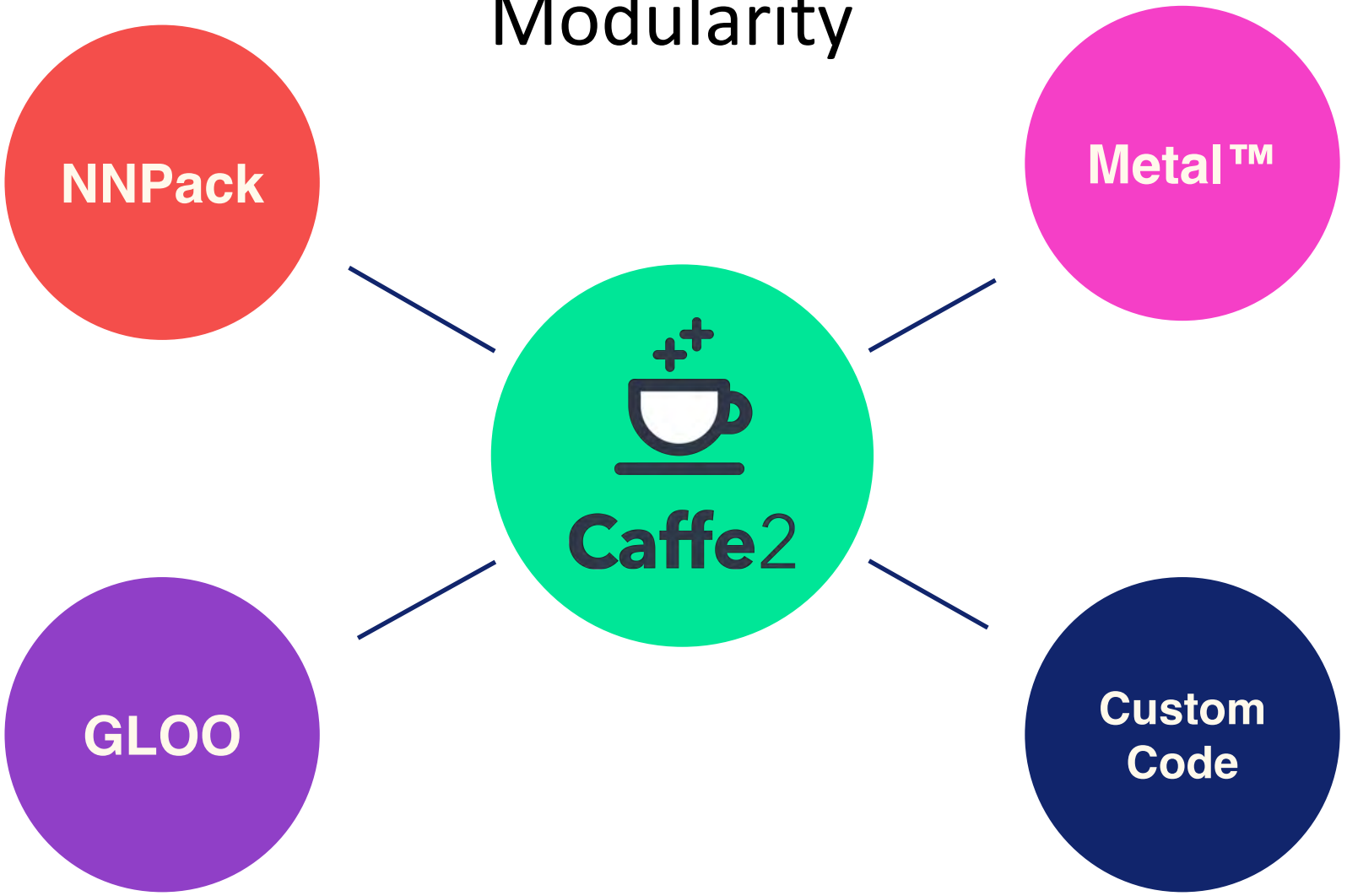
Caffe2 with FP16 Training/Inference

- ~2x training model size & speedups
- Delivering high training throughput on NVIDIA Volta Platforms



Caffe2

Modularity



facebook research


Caffe2
http://caffe2.ai/