Large-Scale Machine Learning at Facebook: Implications of Platform Design on Developer Productivity

Kim Hazelwood
HEAD OF AI INFRA FOUNDATION
FACEBOOK

Mohamed Fawzy
HEAD OF ML TRAINING PLATFORM
FACEBOOK
Facebook Products Leveraging ML

- Search
- Translation
- Face Tagging
- News Feed
- Ads
80% Reduction in Number of Views of Hateful Comments
PORTAL

Smart Camera

Speech Recognition
Computer Vision Advancements

Dense Pose
The Scale of ML at Facebook
The Growth of Data in ML Pipelines

30% FB data used in an ML pipeline in 2018

50% FB data used in an ML pipeline TODAY

2X Data Warehouse Growth since 2018

3X ML Data Growth since 2018
The Growth of ML Training at Facebook

- **Unique Users**: 2X Increase
- **Workflows**: 5X Increase
- **Compute Consumed**: 8X Increase
The Scale of Inference

- 200+ Trillion Total Predictions per Day
- 6.5+ Billion Translations per Day
- 99% Fake Accounts Removed Proactively by Automated Systems Everyday
Where Should We Focus to Enable Better Products?

- Inference
- Training
The Path to Better Products

- BETTER PRODUCTS
- MORE IDEAS
- MORE ENGINEERS
- BETTER EXPERIMENTATION
The Path to Better Experimentation

1.

Ideas are getting more complex

MORE DATA

URU Experiment: 1M Images → 3.5B Images
The Path to Better Experimentation

1. Ideas are getting more complex
2. Models are getting more complex

Complex Models

- GEN 1: DECISION TREES
- GEN 2: SPARSE LINEAR
- GEN 3: NNS: DEEP
- GEN 4: SPARSE

Training Time (days)

Internet-Scale Data / Videos

- 14: IN5K-ResNeXt-101-64x4d
- 7: ResNeXt-101-32x4d
- 4: ResNet-101
- 2: ResNet-50

Months?
The Path to Better Experimentation

1. Ideas are getting more complex
2. Models are getting more complex
3. Iterations per idea is higher

Higher Iterations Per Idea
The Path to Better Experimentation

1. Ideas are getting more complex
2. Models are getting more complex
3. Iterations per idea is higher
4. ML engineer productivity is the secret sauce
How Do We Increase Productivity?

- SW Platforms
- HW Infrastructure
The Case For Infrastructure
Innovation is Driven by Infrastructure Challenges

1. Data & Features
   - Storage Challenges!!

2. Training
   - Network Challenges!!

3. Inference
   - Compute Challenges!!
Training Compute Innovation at Facebook
Scaling Hardware

1. Past
   - HP SL270s (2013)
   - Big Sur (2015)

2. Present
   - Tioga Pass (2017)
     (Dual CPU, High Mem)
   - Big Basin Volta (2018)
     (8X GPU + 2X CPU)

3. Future
   - Zion
     (8 ASICs + 8 CPUs)
Case For Platform

- Experimentation Management
- Auto Tuning Platform
- Distributed Training Platform
  - Support Data & Model Parallel
- Authoring ML Pipelines
- ML Pipeline Runtime Environment
- ML Aware Scheduling & Orchestration

Distributed AI Platform

- Hyper Parameter Tuning, Architecture Search, Feature Selection
- Model Parallel / Data Parallel & Scalability
- Development Lifecycle for ML Pipelines
- Checkpointing & reliable execution. Must have for large scale training
- Abstracts Different Machines (CPUs/GPUs / Accelerators)

FBLearner Flow Platform
End-to-End Developer Productivity
Requires the Collective Vision and Collaboration of All Layers

Model Authoring
Experimentation Platform
Hardware Infrastructure
Performance and Tools

PyTorch
FBLearner
Zion
Timeline View
Innovation Needed at **ALL** Layers

Developer Productivity is the Key to Great Products
Thank you