CSCE 585: Assignment #1
Company: Facebook
Team Members:

V.N. Anirudh Oruganti
Nathan Dolbir
Kennedy Fairey

1-According to Robinson (Source #1), Facebook trains and deploys several hundred models at a time in parallel. Facebook chooses to use models for separate use cases, fine tuning each model using proprietary ML services. Resource-light tasks use simpler models, such as support vector machines (SVM) and logistic regression (LR) being used for face detection and matching. Resource-heavy tasks use a combination of models, namely deep neural networks (DNN) and gradient boosted decision trees (GBDT)

2- Facebook previously used two separate frameworks for model deployment: Caffe2 for production, and PyTorch for research (3). PyTorch was the preferred framework because of its ability for rapid debugging and model parameterization, but they would have to use the ONNX toolchain to transfer PyTorch into production, as it only worked in Python (3). Since then, Caffe2 and PyTorch have merged together into the modern 1.0 and forward version of PyTorch (4), adding Caffe2's focus on cross-platform support and scalability to PyTorch's efficiency. Facebook has also created a proprietary service to assist model training and analysis called FBLearner. It has 3 features: FBLearner Feature Store is used for Facebook's teams to identify and gather useful website data, FBLearner Flow is used to layout the workflow pipeline of ML models, and FBLearner Predictor is used to predict stuff? (6).

3- Facebook does not give insight on load balancing between models, but do have ease in training and implementing models on their backend infrastructure. Big Basin is their cutting-edge GPU server, boasting a 900 GB/s bandwidth and 15.7 teraflops capability (3). Big Basin's throughput makes inter-machine synchronization little to no issue and allows Facebook to scale model training and deployment near-linearly.

4- Facebook does not release details on how their models are retired, but it's suggested that FBLearner Flow may be used for this purpose as it's the ML platform Facebook uses to monitor and implement all of their models (2). Although models might retire mostly likely in the exploration phase where if it underperforms compared to its peers (Source#3).

5- Usually nodes will send their updates via a parameter server where they are aggregated and distributed back to the nodes according to Source #4. The online models at Facebook can be updated every couple of minutes according to Robinson (Source #1).

6- Facebook uses FBLearner to monitor the model pipelines, and uses SGD to continuously update model parameter weights. Same models with different weights are run parallel to continuously reconfigure optimal parameter weights. According to Source#3, models get assigned servers based on the number of parameters, compute and memory usage of the model. The health and performance is monitored in the exploration phase. Models' health/performance are usually checked/compared to their peers based on a number of parameters, accuracy and user feedback via. Views, interactions,etc.

7- Facebook deploys hundreds of models for each of their ML applications and has multiple teams working on models of all sizes. According to Facebook, computer vision models trains on every photo for a few seconds, and speech recognition models train weekly for nearly a day, among other statistics on each given ML application (3)

Service	Resource	Training Frequency	Training Duration
News Feed	Dual-Socket CPUs	Daily	Many Hours
Facer	GPUs + Single-Socket CPUs	Every N Photos	Few Seconds
Lumos	GPUs	Multi-Monthly	Many Hours
Search	Vertical Dependent	Hourly	Few Hours
Language Translation	GPUs	Weekly	Days
Sigma	Dual-Socket CPUs	Sub-Daily	Few Hours
Speech Recognition	GPUs	Weekly	Many Hours

TABLE II
FREQUENCY, DURATION, AND RESOURCES USED BY OFFLINE TRAINING FOR VARIOUS WORKLOADS.

References:

Source #1:

Robinson, Jamal. “How Facebook Scales Artificial Intelligence & Machine Learning.” Medium, Medium, 26 May 2021, jamal-robinson.medium.com/how-facebook-scales-artificial-intelligence-machine-learning-693706ae296f.

Source #2:
https://engineering.fb.com/2016/05/09/core-data/introducing-fblearner-flow-facebook-s-ai-backbone/

Source #3:
https://research.fb.com/wp-content/uploads/2017/12/hpca-2018-facebook.pdf

Source 4:
https://caffe2.ai/blog/2018/05/02/Caffe2_PyTorch_1_0.html

Source 6:
https://www.matroid.com/scaledml/2018/yangqing.pdf

Authors, Various. “Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective.” *Facebook Inc.*, 2018, research.fb.com/wp-content/uploads/2017/12/hpca-2018-facebook.pdf.

Source #8:

https://ai.facebook.com/blog/state-of-the-art-open-source-chatbot/

Source #12:

https://ai.facebook.com/blog/blender-bot-2-an-open-source-chatbot-that-builds-long-term-memory-and-searches-the-internet/