The Non-IID Data Quagmire of Decentralized Machine Learning

Kevin Hsieh, Amar Phanishayee, Onur Mutlu, Phillip Gibbons
ML Training with Decentralized Data

Geo-Distributed Learning

Federated Learning

Data Sovereignty and Privacy
Major Challenges in Decentralized ML

Geo-Distributed Learning

Federated Learning

Challenge 1: Communication Bottlenecks

Solutions: Federated Averaging, Gaia, Deep Gradient Compression
Major Challenges in Decentralized ML

Geo-Distributed Learning

Federated Learning

Challenge 2: Data are often highly skewed (non-iid data)

Solutions: Understudied! Is it a real problem?
Our Work in a Nutshell

Real-World Dataset

Experimental Study

Proposed Solution
Real-World Dataset

Geographical mammal images from Flickr

736K pictures in 42 mammal classes

Highly skewed labels among geographic regions
Skewed data labels are a fundamental and pervasive problem

The problem is even worse for DNNs with batch normalization

The degree of skew determines the difficulty of the problem
Proposed Solution

Replace batch normalization with group normalization

**SkewScout**: communication-efficient decentralized learning over arbitrarily skewed data
Real-World Dataset
Flickr-Mammal Dataset

42 mammal classes from Open Images and ImageNet

40,000 images per class

Clean images with PNAS [Liu et al., '18]

Reverse geocoding to country, subcontinent, and continent

736K Pictures with Labels and Geographic Information

https://doi.org/10.5281/zenodo.3676081
Top-3 Mammals in Each Continent

Each top-3 mammal takes 44-92% share of global images
Label Distribution Across Continents

Vast majority of mammals are dominated by 2-3 continents

The labels are even more skewed among subcontinents
Experimental Study
Scope of Experimental Study

**ML Application**
- Image Classification (with various DNNs and datasets)
- Face recognition

**Decentralized Learning Algorithms**
- Gaia [NSDI’17]
- FederatedAveraging [AISTATS’17]
- DeepGradientCompression [ICLR’18]

**Skewness of Data Label Partitions**
- 2-5 Partitions -- more partitions are worse
Results:

- GoogLeNet over CIFAR

- Shuffled Data
  - Top-1 Validation Accuracy:
    - BSP (Bulk Synchronous Parallel): 80%
    - Gaia (20X faster than BSP): 68%
    - FederatedAveraging (20X faster than BSP): 71%
    - DeepGradientCompression (30X faster than BSP): 40%

- Skewed Data
  - Top-1 Validation Accuracy:
    - BSP (Bulk Synchronous Parallel): 20%
    - Gaia (20X faster than BSP): -12%
    - FederatedAveraging (20X faster than BSP): -15%
    - DeepGradientCompression (30X faster than BSP): -69%

All decentralized learning algorithms lose significant accuracy.

Tight synchronization (BSP) is accurate but too slow.
Skewed data is a pervasive and fundamental problem

Even BSP loses accuracy for DNNs with Batch Normalization layers
Degree of Skew is a Key Factor

Degree of skew can determine the difficulty of the problem
Batch Normalization — Problem and Solution
Background: Batch Normalization

[ioffe & Szegedy, 2015]

Standard normal distribution \((\mu = 0, \sigma = 1)\) in each minibatch at training time

Batch normalization enables larger learning rates and avoid sharp local minimum (generalize better)
Batch Normalization with Skewed Data

Minibatch Mean Divergence: \[ \frac{|\text{Mean}_1 - \text{Mean}_2|}{\text{AVG}(\text{Mean}_1, \text{Mean}_2)} \]

Minibatch mean and standard deviation vary significantly among partitions.

Global \( \mu \) and \( \sigma \) do not work for all partitions.

CIFAR-10 with BN-LeNet (2 Partitions)
Solution: Use Group Normalization [Wu and He, ECCV’18]

Batch Normalization

Group Normalization

Designed for small minibatches
We apply as a solution for skewed data
**Results with Group Normalization**

GroupNorm recovers the accuracy loss for BSP and reduces accuracy losses for decentralized algorithms.
SkewScout: Decentralized learning over arbitrarily skewed data
Overview of **SkewScout**

- Recall that **degree of data skew** determines **difficulty**.
- **SkewScout**: **Adapts** communication to the **skew-induced accuracy loss**.

Minimize commutation when accuracy loss is acceptable. Work with different decentralized learning algorithms.
Evaluation of SkewScout

All data points achieve the same validation accuracy

Significant saving over BSP
Only within 1.5X more than Oracle

CIFAR-10 with AlexNet
CIFAR-10 with GoogLeNet
Key Takeaways

• **Flickr-Mammal dataset**: Highly skewed label distribution in the real world

• Skewed data is a pervasive problem
• Batch normalization is particularly problematic

• **SkewScout**: adapts decentralized learning over arbitrarily skewed data
• Group normalization is a good alternative to batch normalization