DeepMind

One Pass ImageNet

Huiyi Hu, Ang Li, Daniele Calandriello, Dilan Gorur

In a nutshell

Traditional multi-epoch setup:

- 90-epoch ResNet50: 76% top-1 accuracy
- Not efficient and require full access to all data

The One Pass Imagenet (OPIN) benchmark:
how well can we learn in a single epoch/pass of Imagenet with limited replay memory?

Motivation

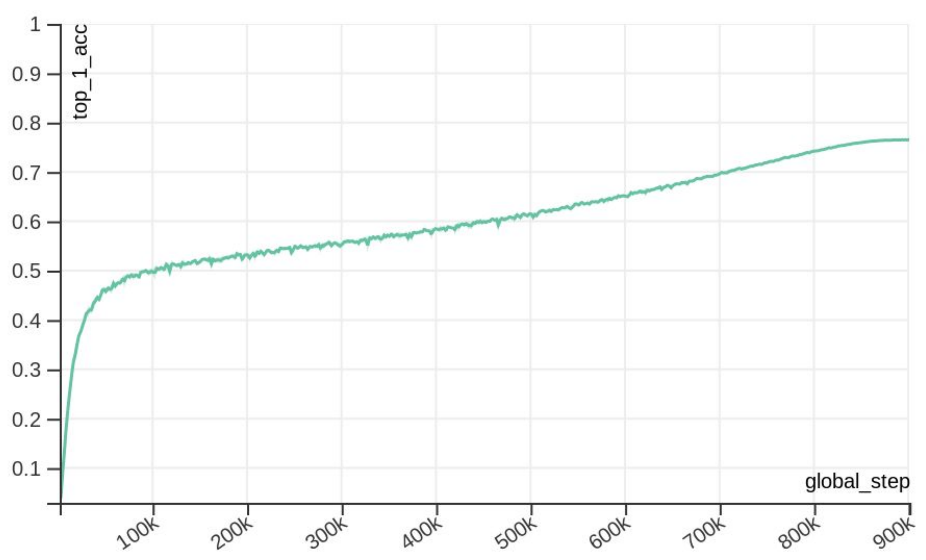
- Streaming data is common in real world with constraints on storage.
- Real world data is large scale. Imagenet is a first step towards that.
- Study accuracy/memory/compute trade-off
- One pass has implicit non-stationarity

	Accuracy (%) ↑	Storage (%) ↓	Compute (%) ↓
Multi-epoch (90 epochs)	76.9	100	100
One-Pass (Naive)	30.6	0	1.1
One-Pass (Prioritized Replay)	65.0	10	10

Problem Setup

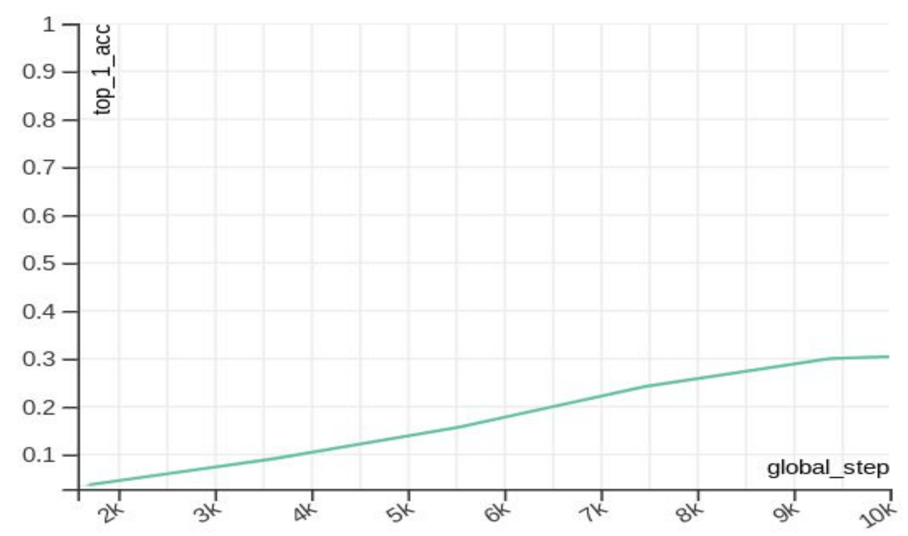
- One pass of Imagenet (train from random initialization)
- Multi-metrics: top-1 accuracy, replay memory, compute
- Random ordering of data

North star:
90-epoch training



90 epochs for 76% accuracy

Naive baseline:
1-epoch training (without replay)



1 epoch for 30% accuracy

Ingredients of a baseline

Replay buffer

- Backend: Reverb server
- k replay steps with augmentation (effectively k+1 epoch of compute)

Priority Sampling

- Uniform is very hard to beat!
- Error-based Priority Replay (EPR) with decay schedule

$$P(x,y) = 1 - \alpha e^{-\ell(x,y;\theta)}$$

Importance Weight

- Priority sampling from Replay changes its data distribution from p(x) to q(x)
- To compensate for the change such that

$$\mathbb{E}_q[w(x)\ell(x;\theta)] = \mathbb{E}_p[\ell(x;\theta)]$$

- we **re-weight** the loss on replay samples with:

$$w(x) \propto 1/Priority(x)$$

Experiments

- Trade-offs of accuracy / compute / memory
- Uniform replay sampling is a strong baseline. Priority replay brings small improvement:
  64.7% -> 65.0% with std 0.07%
- Performance saturates as compute/ memory increase. Still room for improving the utilization of extra compute/ memory.
- Comparison to regular multi-epoch performance at 2, 4, 6, 9 epochs.

Effective Epochs	Computation	Storage (Prioritized Replay) 1 %	5 %	10 %	Multi-epoch 100% Storage
2	$2/90 \approx 2.2\%$	44.7	45.1	45.7	46.1
4	$4/90 \approx 4.4\%$	55.5	57.1	57.2	59.0
6	$6/90 \approx 6.7\%$	58.9	61.3	62.2	64.1
9	$9/90 = 10\%$	59.3	63.2	65.0	68.2

Open questions

- Introduce explicit distributional shift?
- One pass learning starting from a pre-trained model?
- How to better utilize the extra memory and compute?
- What's a good sampling scheme for efficient learning?