High-Efficiency Online Data Generation to Improve Pretraining Scaling Laws of Deep Networks

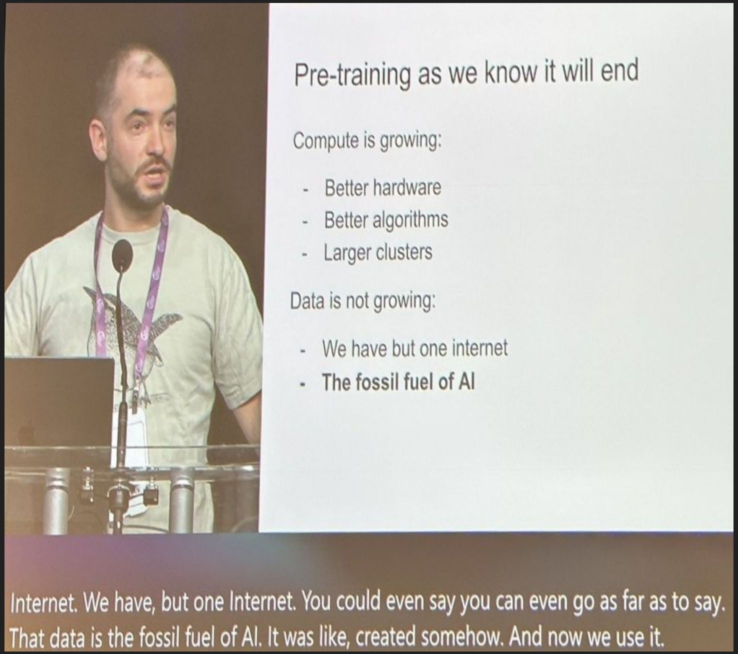
Sike Ogieva, Sushmit Chakma
Advised by: Dr. Randall Balestriero

Research Question

How can we integrate on-the-fly data generation with reinforcement learning to optimize neural network training in data-constrained environments?

- **Goal:** Build a system that generates data in real-time, tailored to the model's current weaknesses, optimizing training.
- **Novelty:** Merges efficient on-the-fly data generation with reinforcement learning's adaptive decision-making—unexplored territory.

Motivation: The Data Dilemma in Machine Learning



Challenges:

- Data scarcity and high costs in acquiring real data.
- Privacy and regulatory restrictions limit data access.
- Vast storage requirements hinder training with large datasets.

Impact:

- Example: Facebook trained face recognition on ~4.4 million images.
- Many industries (healthcare, finance, autonomous vehicles) face similar hurdles.

Motivation: Automatic Data Augmentation

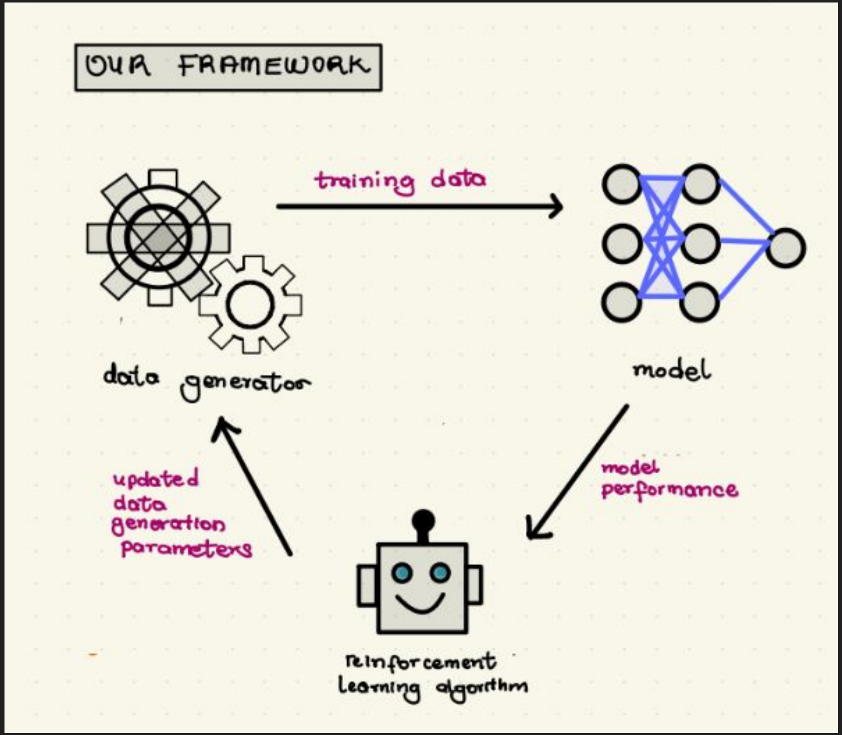
Dataset	GPU hours	Best published results	Our results
CIFAR-10	5000	2.1	1.5
CIFAR-100	0	12.2	10.7
SVHN	1000	1.3	1.0
Stanford Cars	0	5.9	5.2
ImageNet	15000	3.9	3.5

Table 1. Error rates (%) from this paper compared to the best results so far on five datasets (Top-5 for ImageNet, Top-1 for the others). Previous best result on Stanford Cars fine-tuned weights originally trained on a larger dataset [66], whereas we use a randomly initialized network. Previous best results on other datasets only include models that were not trained on additional data, for a single evaluation (without ensembling). See Tables 2,3, and 4 for more detailed comparison. GPU hours are estimated for an NVIDIA Tesla P100.

Selecting data augmentation instances that conform with a learning model's apparent weaknesses is a powerful way to improve model performance.

Cubuk et al, AutoAugment: Learning Augmentation Policies from Data, 2019

Proposed Methodology



Dynamic Data Generation:

- Generate data in batches on demand using the OTF framework.

Minimal Storage:

- Store only essential parameters (e.g., 21,600 parameters per day vs. 86,400 raw values).

RL-Driven Optimization:

- An RL agent observes current model performance (state) and selects the next data generation action to maximize improvements (reward).
- This forms a Markov Decision Process (MDP) guiding the cycle.

Literature Review

1. "On the Fly" Framework
   (Vejdan et al., 2019)

2. Infigen: Pure Procedural Photorealistic Image Generation
   (Raistrick et al, 2023)

3. Cheaper, Faster Automatic Data Augmentation
   (Li et al, 2020; Lim et al, 2020)

4. On-the-fly Dataset Augmentation with Synthetic Data
   (Li et al, 2025)

5. Procedural Image Generation Resources
   https://procgen.space/resources
6. ...

Next Up

First prototype of framework

1. Data Generator
   a. Write procedural MNIST and Fashion-MNIST generators
   b. Use procedural libraries (like firesim and infigen) on relevant classification problems
2. Image Classifier
   a. Small convolutional model
3. Reinforcement Learning Algorithm
   a. AutoAugment, Fast AutoAugment and Differentiable AutoAugment

Then evaluation and reiteration with bigger generators and a bigger neural net...

Applications

1. **Healthcare**
   Generate synthetic patient records for training diagnostic models while preserving privacy.

2. **Finance**
   Create synthetic transaction data for fraud detection and risk modeling.
   Fraud detection, in particular, has a data imbalance problem

3. **Autonomous Vehicles**
   Produce real-time simulated driving scenarios for training under rare conditions.

And More...