MNIST the Hard Way

Comparing Evolution Strategies and Stochastic Gradient Descent

William Saunders

Department of Computer Science
University of Toronto

Jan 26, 2018

Sources

On the Relationship Between the OpenAI Evolution Strategy and Stochastic Gradient Descent
By: Xingwen Zhang, Jeff Clune, Kenneth O. Stanley (Uber AI labs)

OpenAI Evolution Strategies

- Generate n pseudo-offspring by adding perturbation v_i sampled from uniform Gaussian with mean 0, covariance I $\theta + \sigma v_i$
- Compute rewards for each offspring (r_i)
- Estimate gradient

$$g^{ES} = \frac{1}{n\sigma} \sum_{i=1}^{n} v_i r_i$$

- Update parameters using gradient estimate with any optimizer (ie. ADAM)

Evolution Strategies vs Finite-difference

- Finite-difference gradient approximation:

$$\frac{\partial}{\partial \theta_i} f(\theta) \approx \frac{f(\theta + h * e_i) - f(\theta)}{h}$$

  where e_i is 1 at index i, 0 everywhere else
- d function evaluations, where d is dimension of θ
- Evolution strategies approximation:

$$\frac{\partial}{\partial \theta_i} f(\theta) \approx \mathop{\mathbb{E}}_{v \sim N(0,I)} \frac{v_i f(\theta + \sigma v)}{\sigma}$$

  Since: $\mathbb{E}_{v \sim N(0,I)} \frac{v_i f(\theta)}{\sigma} = 0$

$$\frac{\partial}{\partial \theta_i} f(\theta) \approx \mathop{\mathbb{E}}_{v \sim N(0,I)} \frac{v_i (f(\theta + \sigma v) - f(\theta))}{\sigma}$$

- n function evaluations, independent of dimension of θ

Setup

- We can use ES gradient as an inefficient way to train a classification network on MNIST, to compare it's performance to the SGD gradient
- We can compute the correlation between the SGD gradient and ES gradient for each minibatch
- Initially achieved 96.99% validation accuracy with 10,000 pseudo-offspring, 2000 iterations on a network with 3,274,634 parameters

Gradient Correlation

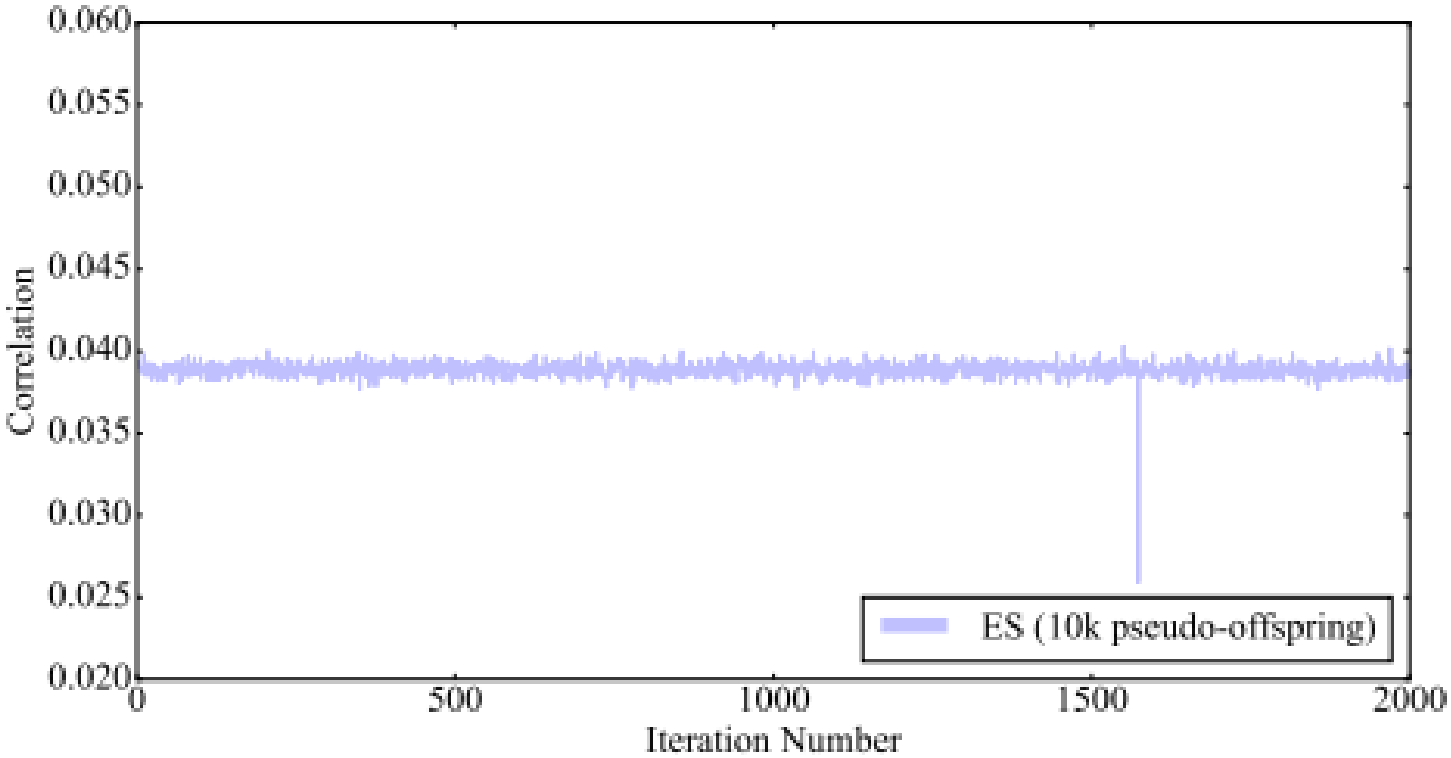

Figure 1: **Correlation of the gradients estimated by ES and the analytic gradients for the same sequence of mini-batches.** The correlation between ES and gradients used by SGD is remarkably stable.

Noisy SGD Proxy

- It's remarkable that the 3.9% correlation allows for validation accuracy within 1.7% of SGD accuracy
- The correlation between ES and SGD gradient is stable over time
- We can simulate ES gradient by adding uniform noise to SGD gradient to achieve the same correlation to the SGD gradient
- It's more efficient to run experiments on algorithm changes with this SGD + noise proxy

Noisy SGD Proxy

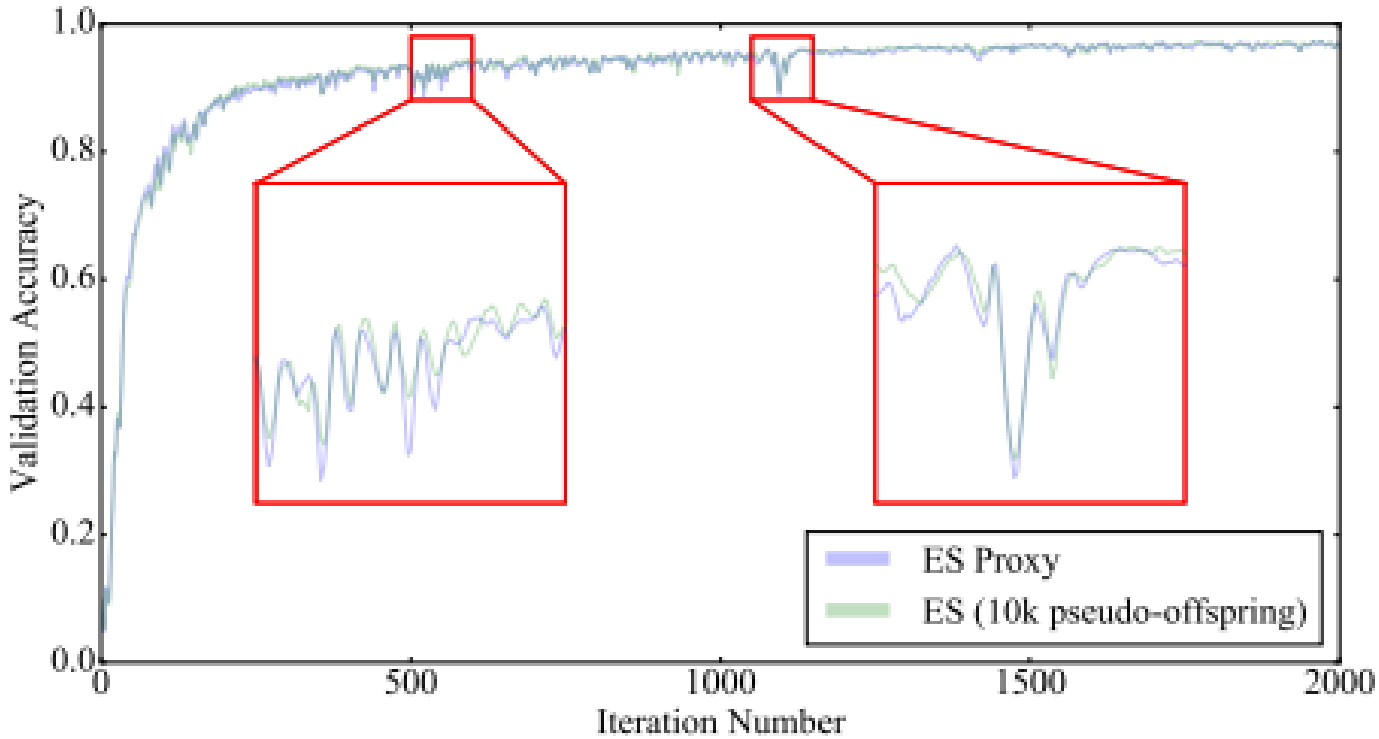

Figure 2: **Validation accuracy of ES and the ES proxy for the same sequence of mini-batches.** Notice how the fluctuations in performance by ES and its proxy are nearly identical throughout the run despite the randomness in the SGD proxy and the ES pseudo-offspring. Insets show local areas of the curves close up. The implication is that mini-batches are the primary driver of noise in the search, impacting ES and SGD in the same way. Recall that validation accuracy over iterations in this figure and throughout this paper is reported on the entire MNIST test set.

MNIST The Hard Way

- Limited the number of dimensions that are perturbed for each pseudo-offspring, improving performance but requiring more pseudo-offspring
- Achieve 99% accuracy on MNIST using 50,000 pseudo-offspring, 10,000 training batches
- But used a smaller network (2 layers, 28938 parameters)
- Used approximately enough pseudo-offspring to perform finite differences!

Summary

- Evolution Strategies with fixed σ performs similarly to gradient descent plus noise, but is much more expensive

Improve ES stability - No-mini-batch ES