How Much More Data Do I Need? Estimating Requirements for Downstream Tasks

Rafid Mahmood, James Lucas, David Acuna, Daiqing Li, Jonah Philion, Jose M. Alvarez, Zhiding Yu, Sanja Fidler, Marc T. Law

The Data Collection Problem

We have a model, initial data n_0, & target validation score V^*. We can order data over T rounds, but we want to collect the least amount needed to meet the target.

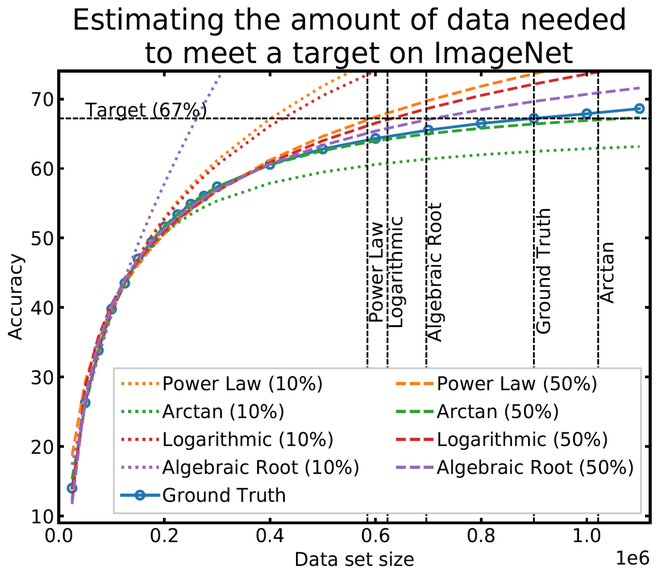
In each round:

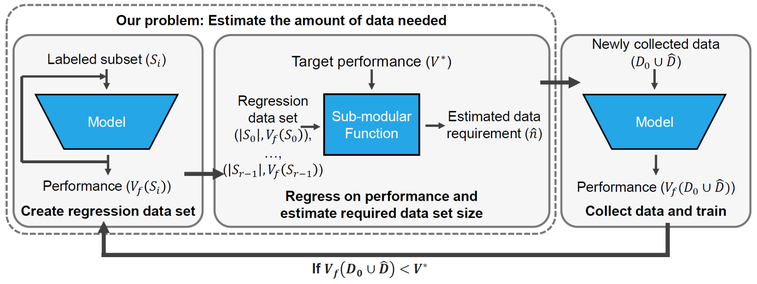
1. Estimate how much more data we need $\hat{n}$.
2. Sample data until we have $\hat{n}$ points, train the model, & evaluate the score.

Initial Approach

Fit neural scaling law functions on a small subset of data, extrapolate learning curve, solve for minimum data needed.

- **If we start with a lot of data (50% of data), any function can fit learning curves.**
- **With small data sets (10%), all functions diverge when extrapolating.**
- **Good extrapolation (6% error) can still give poor data estimate (300,000 fewer images).**

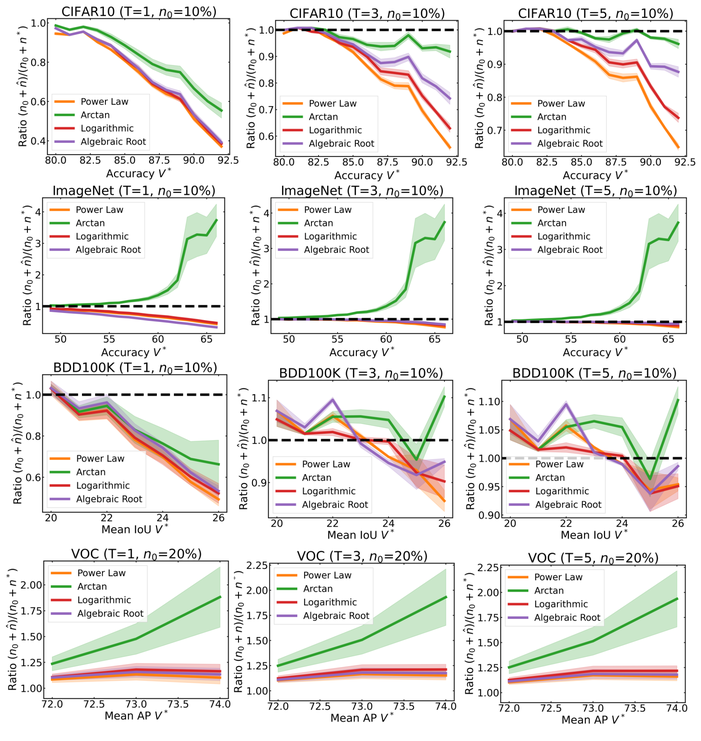




Experiments

Fit regression function with initially n_0% of the data & simulate for T rounds. Evaluate the ratio $\frac{n_0+\hat{n}}{n_0+n^*}$, where $\hat{n}$ is the total collected & and n^* is the minimum amount needed.

- **Most regression functions significantly over- or under-estimate how much data we need.**

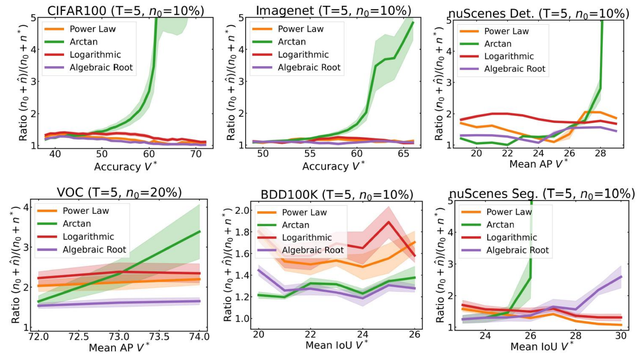


Regression Function	Equation	Predicting how much data is needed to meet the target
Power Law	$v(n;\theta) = \theta_1 n^{\theta_2} + \theta_3$	Under-estimates
Arctan	$v(n;\theta) = \frac{200}{\pi}\arctan\left(\theta_1 \frac{\pi}{2} n + \theta_2\right) + \theta_3$	Over-estimates
Algebraic Root	$v(n;\theta) = \frac{100n}{(1+\lvert\theta_1 n\rvert^{\theta_2})^{\frac{1}{\theta_2}}} + \theta_3$	Under-estimates
Logarithmic	$v(n;\theta) = \theta_1 \log(n+\theta_2) + \theta_3$	Under-estimates

Using a Correction Factor

Take a previous task (e.g., CIFAR10) & solve for τ such that if we target for $V^* + \tau$, we will always meet V^*.

- **Functions that used to under-estimate now collect enough data.**



Insights & Practical Guidelines

- Use T ≈ 5 rounds of data collection with techniques that under-estimate.
- Use past tasks to identify which functions under-estimate & learn a correction factor. With 5 rounds & correction factor, we collect at most 1.5× the minimum data.
- Use all the regression functions to obtain an interval that approximately bounds the true data requirement.