Internet Curiosity: Directed Learning on Uncurated Internet Data

Alexander C. Li*[1], Ellis Brown*[1], Alexei A. Efros[2], Deepak Pathak[1]
[1]Carnegie Mellon University, [2]UC Berkeley, *equal contribution

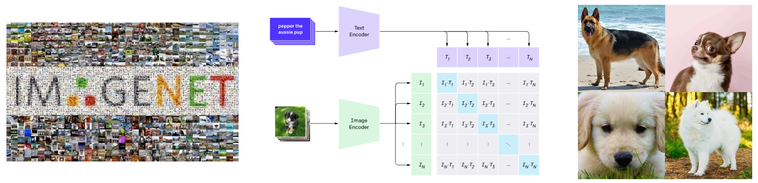
ECCV TEL AVIV 2022

Carnegie Mellon University

BAIR BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Background and Motivation

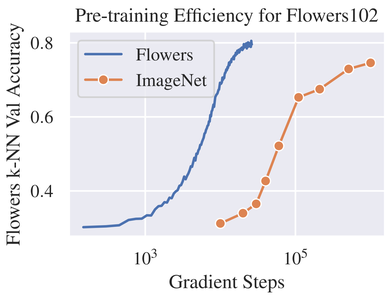
Standard Transfer Learning Setup



1. Curate diverse dataset 2. Pre-train 3. Fine-tune on target

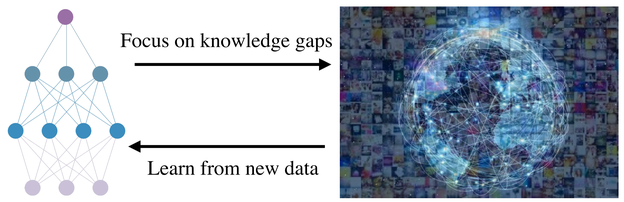
- Only data relevant to the target task improves model generalization, while everything else wastes time and compute!

Case Study: Efficient Pre-training on Flowers102



	Pre-training Dataset	
	Flowers	ImageNet
Train size	2040	1.28M
Steps	25k	500k
Flowers Acc.	**79.6%**	74.6%

Solution: Open-world Learning on the Internet!



Two main challenges:

1. *How do we handle the trillions of photos on the Internet?*
2. *Which images are relevant to the target task?*

Our solution:

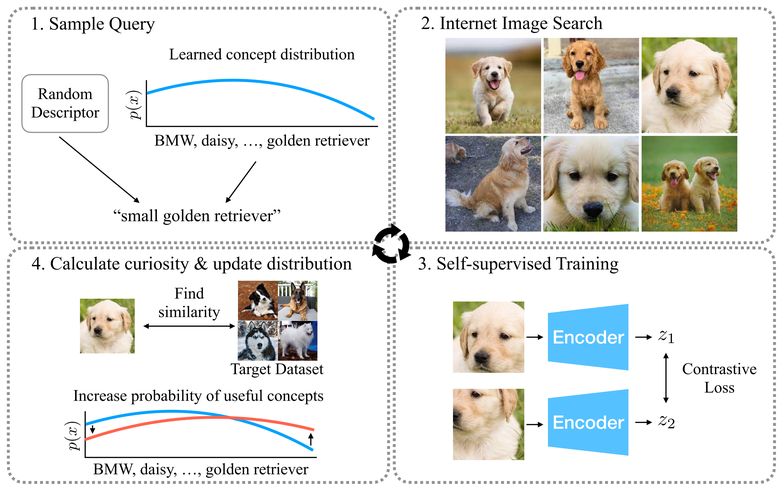
- Query Google Image Search to make exploring the Internet easier.
- Self-supervised image score (similarity to target images).
- Use text similarity to predict what unseen queries are helpful.
- Self-supervised training on downloaded images.

Internet Curiosity Method

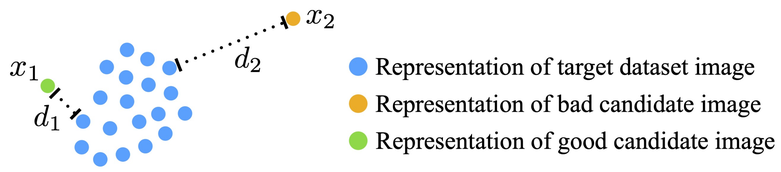
Settings

- Self-supervised: only have unlabeled images from the target task.
- Semi-supervised: have unlabeled images from the target task, as well as the label set (e.g., {'husky,' 'chihuahua,' 'poodle,' ...})

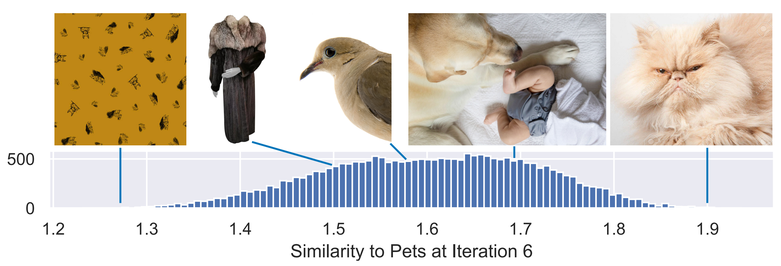
Method Overview



Self-supervised Image Relevance Score

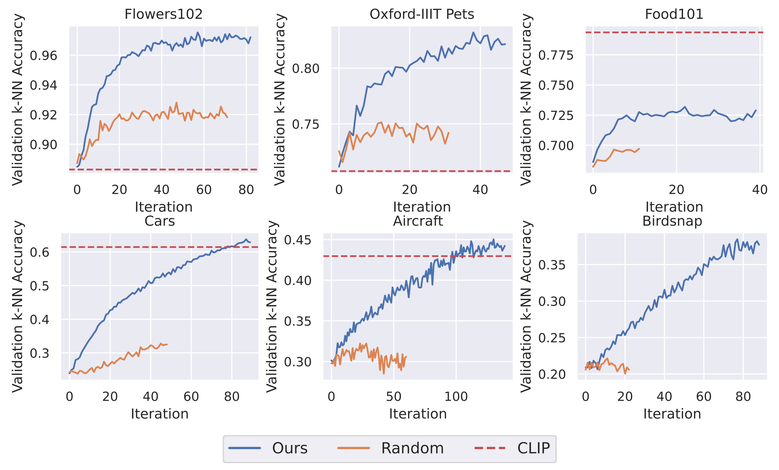


- Reward for a candidate representation x_i is $-d_{\cos}(x_i, y_j)$, where y_j is its nearest neighbor in the target representations.



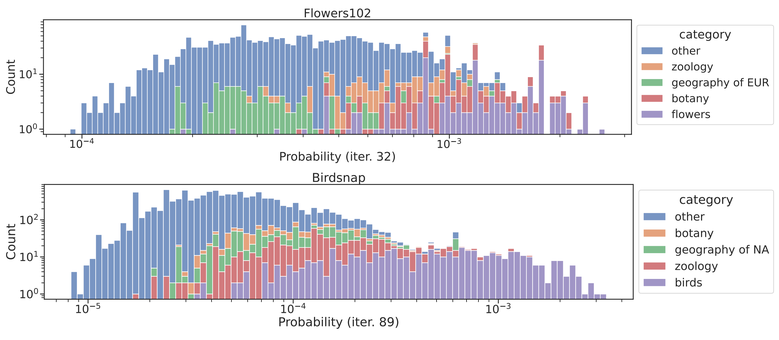
Results

Britannica Encyclopedia Corpus (no knowledge of label set)



- Vocabularies are 5% class labels, 95% random Britannica nouns

Discovered Category Probabilities



Wordnet Corpus (with knowledge of label set)