Efficient Learning

Danna Gurari
University of Colorado Boulder
Fall 2023

Review

- Last lecture on model compression:
  - Motivation
  - Key idea: knowledge distillation (KD)
  - Pioneering KD model for image classification
  - Pioneering KD model for object detection
  - State-of-the-art for KD (ICCV 2023 highlights)

- Assignments (Canvas):
  - Project presentation (poster and video) due in 1.5 weeks
  - Project report due in 2 weeks

- Questions?

Efficient Learning: Today's Topics

- Motivation
- Curriculum Learning
- Active Learning
- Few-shot Learning
- Faculty Course Questionnaire (FCQ)

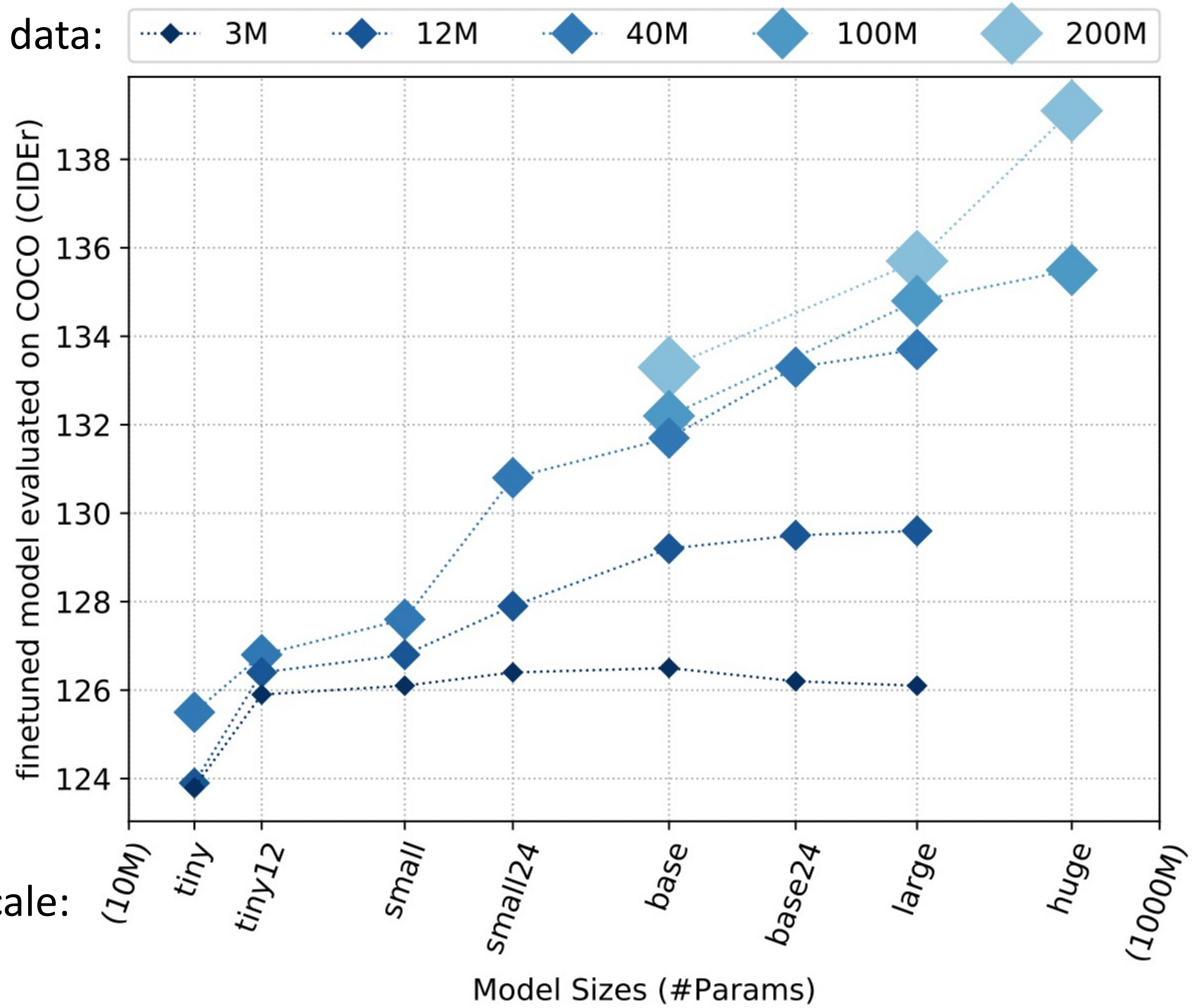
Efficient Learning: Today’s Topics

- Motivation
- Curriculum Learning
- Active Learning
- Few-shot Learning
- Faculty Course Questionnaire (FCQ)

Trend: Extensive Training

Models perform better with more training data (paired with parameter-heavy architectures):

Amount of training data:



Logarithmic scale:

Hu et al. Scaling Up Vision-Language Pre-training for Image Captioning. CVPR 2022

Trend: Extensive Training

How many training examples lead to top performance in Vision Transformers?

- 3 million
- 30 million
- 300 million
- 3 billion
- 30 billion

It takes 2,500 TPUv3- core-days to train this model

Zhai et al. Scaling Vision Transformers. CVPR 2022

Boss: What did you do last month?

You: Trained the model for one epoch.

Boss: Umm, fine, what is your plan for next month?

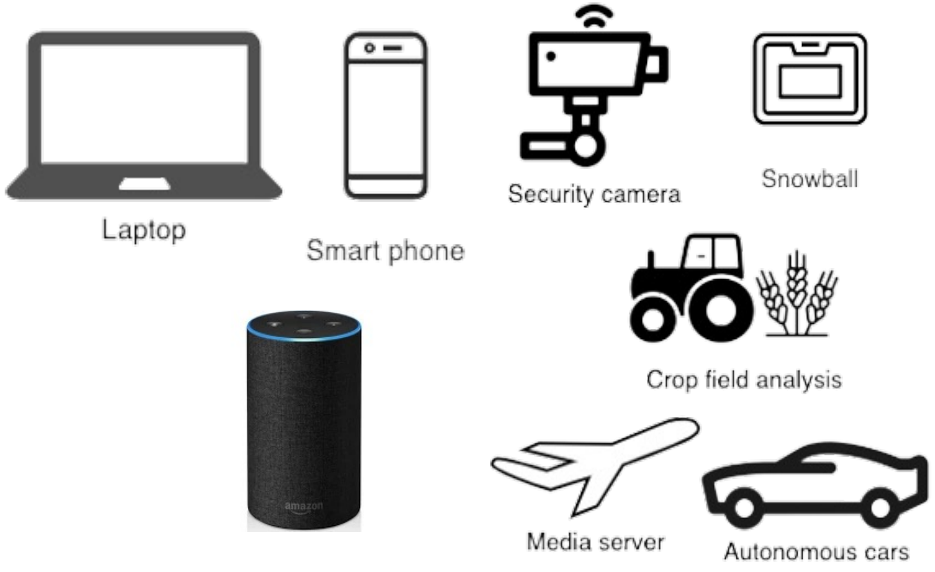
You: Train… train the model for one more epoch?

https://hanlab.mit.edu/files/course/slides/MIT-TinyML-Lec13-Distributed-Training-I.pdf

Why Is Extensive Training Costly?

- Time-consuming
- Expensive
- Increased environmental impact from carbon emissions

When Is Extensive Training Unrealistic?



1. On-device adaptation (e.g., because of privacy concerns and poor/no internet connection):

2. Rare content for which there is a scarcity of data (e.g., private content including medical information, natural disasters, rare locations such as outer space)

Figure: https://aws.amazon.com/blogs/machine-learning/demystifying-machine-learning-at-the-edge-through-real-use-cases/

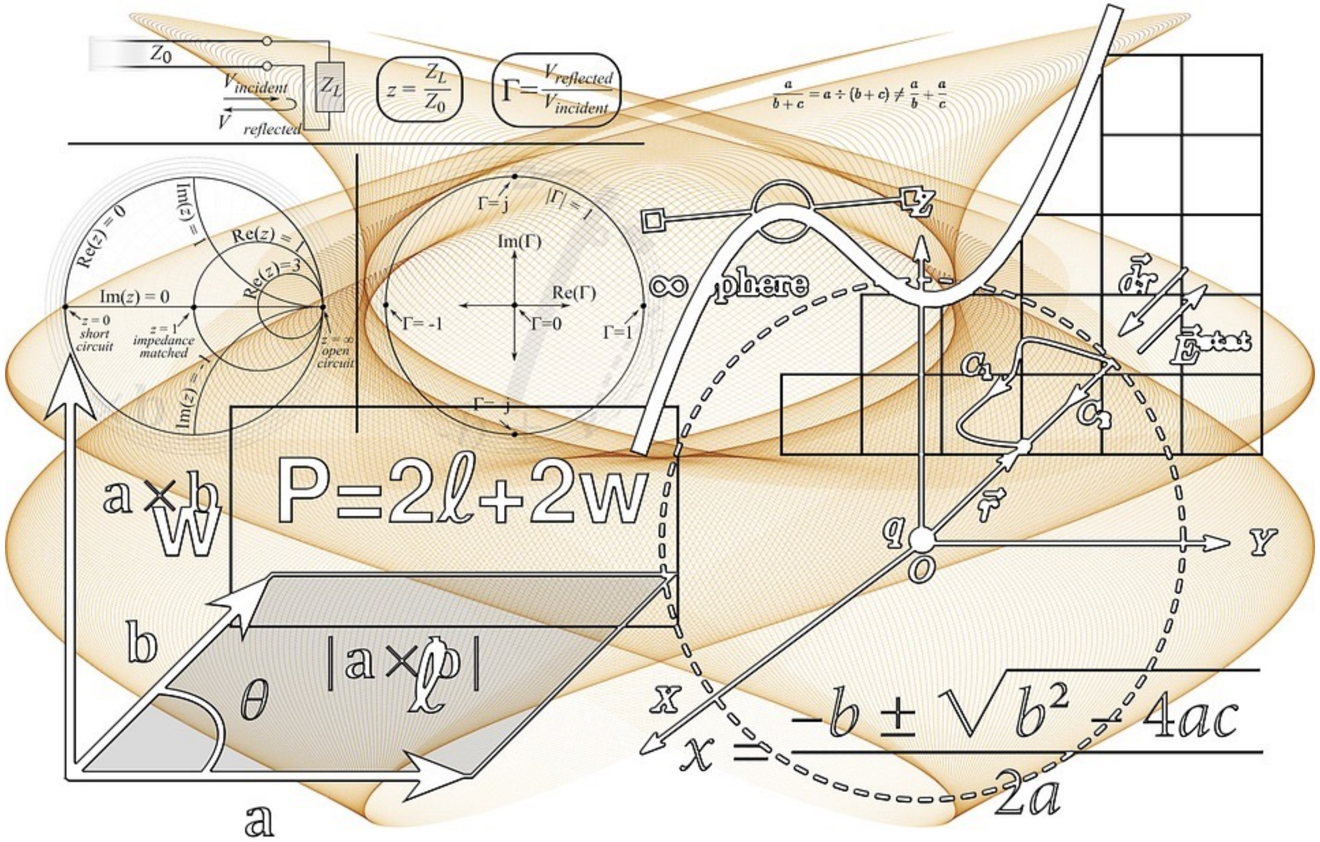
How to teach machines so they learn more efficiently: (1) faster and (2) with fewer resources?

Efficient Learning: Today’s Topics

- Motivation
- Curriculum Learning
- Active Learning
- Few-shot Learning
- Faculty Course Questionnaire (FCQ)

Intuition: How to Teach a Child Math?

Random Order of Examples



Meaningful Order of Examples


Table of Contents

Big Book of Math; Dinah Zike

Intuition: How to Teach a Child To Read

Random Order of Examples

Meaningful Order of Examples

Idea: Teach Machines As We Teach Humans

Curriculum

Train with simpler examples first and progressively harder examples over time

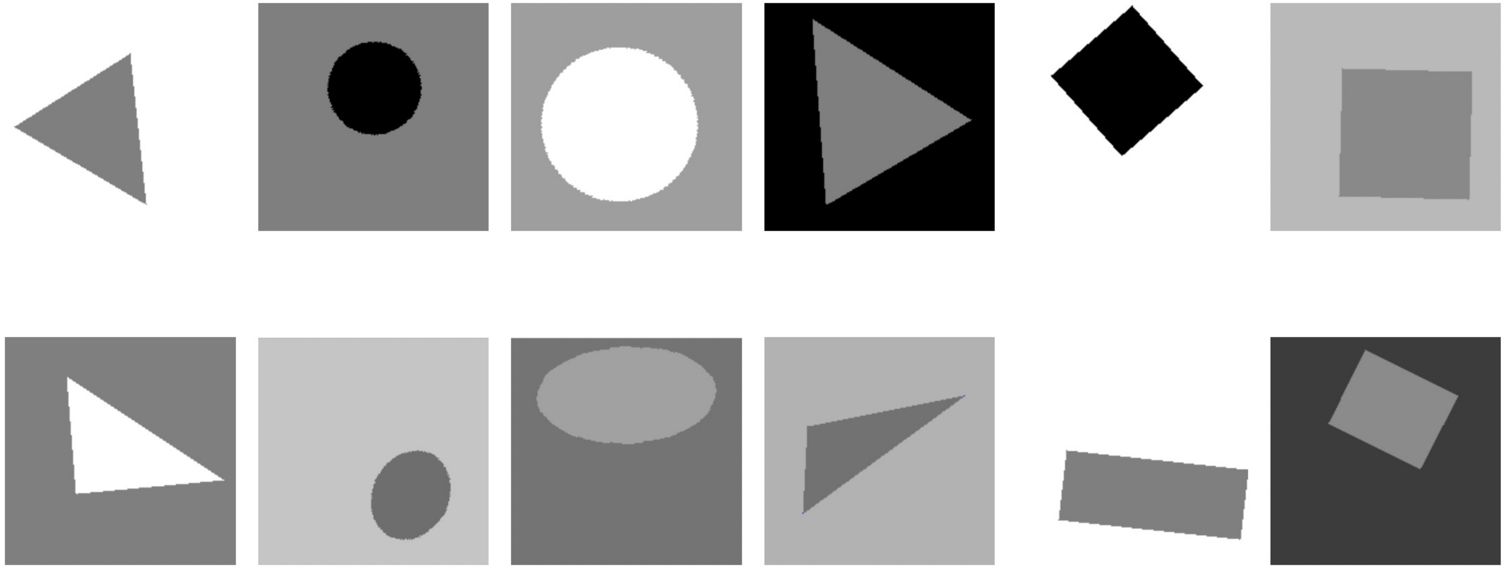
Jeffrey L. Elman. Learning and development in neural networks: The importance of starting small. Cognition, 1993.

Key Evaluation Metrics

- Training convergence speed
- Generalization performance on test data

Pioneering Task: Shape Prediction

Classify each shape as rectangle, ellipse, or triangle



Solution: 3-layer neural network

1. Easy (Basic): less shape variability (squares, circles, and equilateral triangles); 10,000 examples
2. Hard (Geom): more shape variability (rectangles, ellipses, and triangles); 10,000 examples

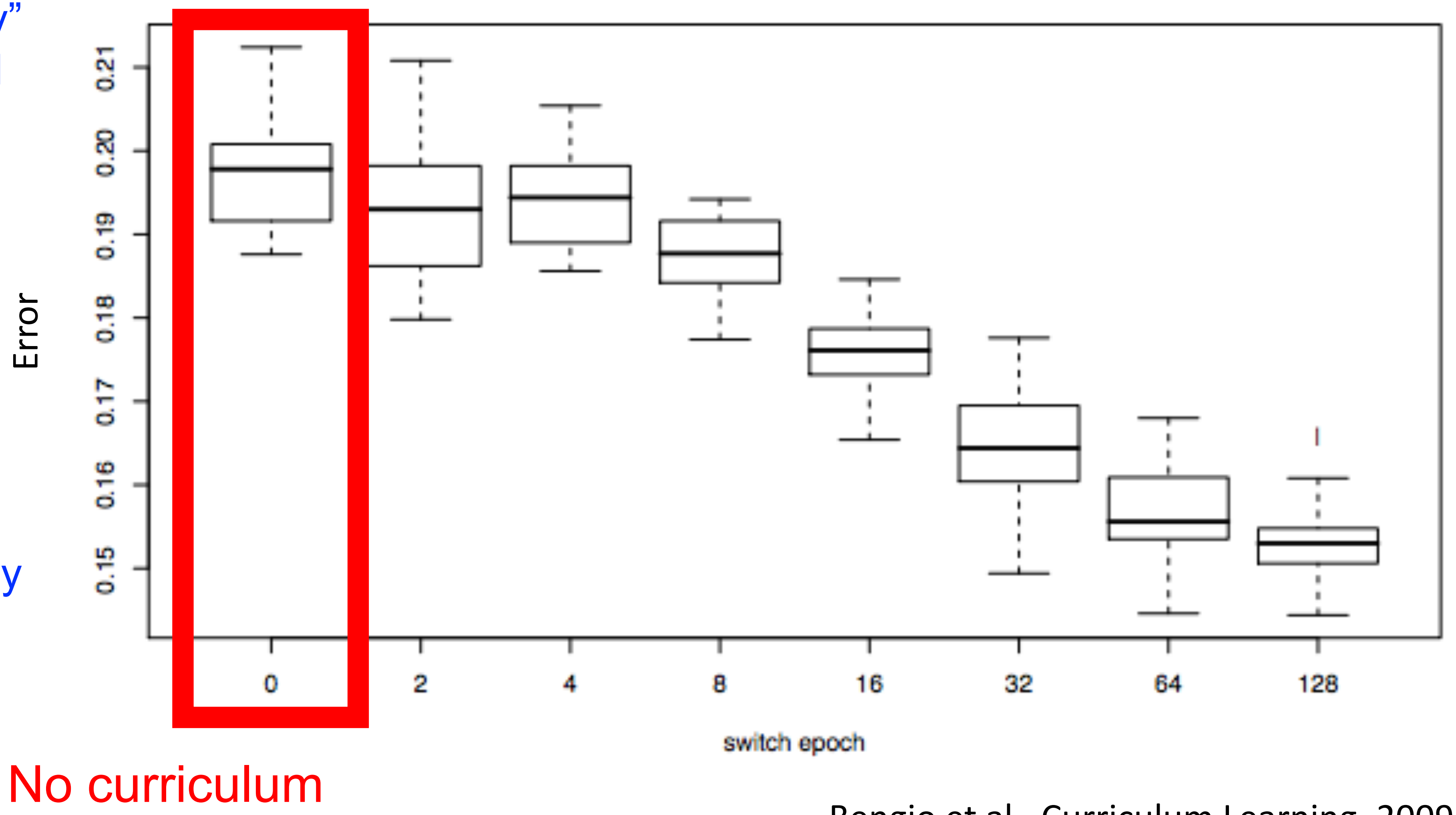
Bengio et al., Curriculum Learning, 2009

Shape Prediction: Curriculum Learning

Results of training on “easy” examples for *n* epochs and then training on “hard” examples until 256 epochs (20 random initializations).

What are benefits of curriculum learning?

How many epochs should the algorithm train with easy examples before switching to difficult examples?



Bengio et al., Curriculum Learning, 2009

EfficientTrain: An ICCV 2023 Paper

EfficientTrain: Exploring Generalized Curriculum Learning for Training Visual Backbones

Yulin Wang[1*] Yang Yue[1*] Rui Lu[1] Tianjiao Liu[2] Zhao Zhong[2]
Shiji Song[1] Gao Huang[1,3✉]
[1]Department of Automation, BNRist, Tsinghua University [2]Huawei Technologies Ltd. [3]BAAI
{wang-yl19, yueyang22}@mails.tsinghua.edu.cn, gaohuang@tsinghua.edu.cn

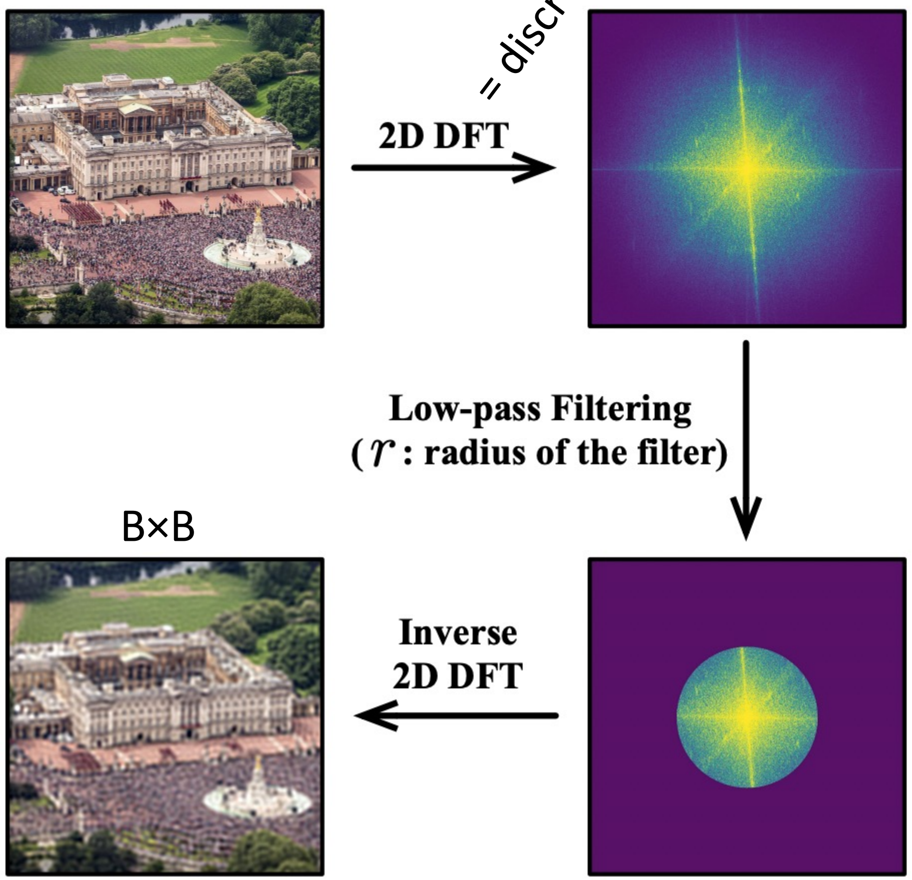
Key idea: eliminate difficult patterns from all training examples at earlier learning stages by removing higher-frequency content

EfficientTrain: Key Idea

~20% training cost eliminated by initially training on lower resolution, low-frequency images to learn low-frequency information typically learned first during training

H×W

= discrete Fourier transform



B×B

(a) Low-pass Filtering
(DFT: discrete Fourier transform)

B×B patch cropped in frequency domain

Recent Work: Another ICCV 2023 Paper

Learning to Learn: How to Continuously Teach Humans and Machines

Parantak Singh[1,2], You Li[2,3], Ankur Sikarwar[1,2], Weixian Lei[4], Difei Gao[4], Morgan B. Talbot[5,6], Ying Sun[2], Mike Zheng Shou[4], Gabriel Kreiman[5], Mengmi Zhang[1,2]

[1] Nanyang Technological University (NTU), Singapore [2] CFAR and I2R, Agency for Science, Technology and Research, Singapore, [3] University of Wisconsin-Madison, USA, [4] Show Lab, National University of Singapore, Singapore, [5] Boston Children's Hospital, Harvard Medical School, USA, [6] Harvard-MIT Health Sciences and Technology, MIT,

Address correspondence to mengmi@i2r.a-star.edu.sg

First to study CL for online class-incremental learning, meaning the curriculum orders different classification classes/tasks and permits observing each training example once

- Key challenge: avoid forgetting previously learned tasks

Key Questions In Creating “Curriculum”

- How to define what is “easy” versus “hard”?

- How many levels to include in the curriculum from easy to hard?

Breakout rooms: choose a task and then address the above two questions

Efficient Learning: Today’s Topics

- Motivation
- Curriculum Learning
- Active Learning
- Few-shot Learning
- Faculty Course Questionnaire (FCQ)

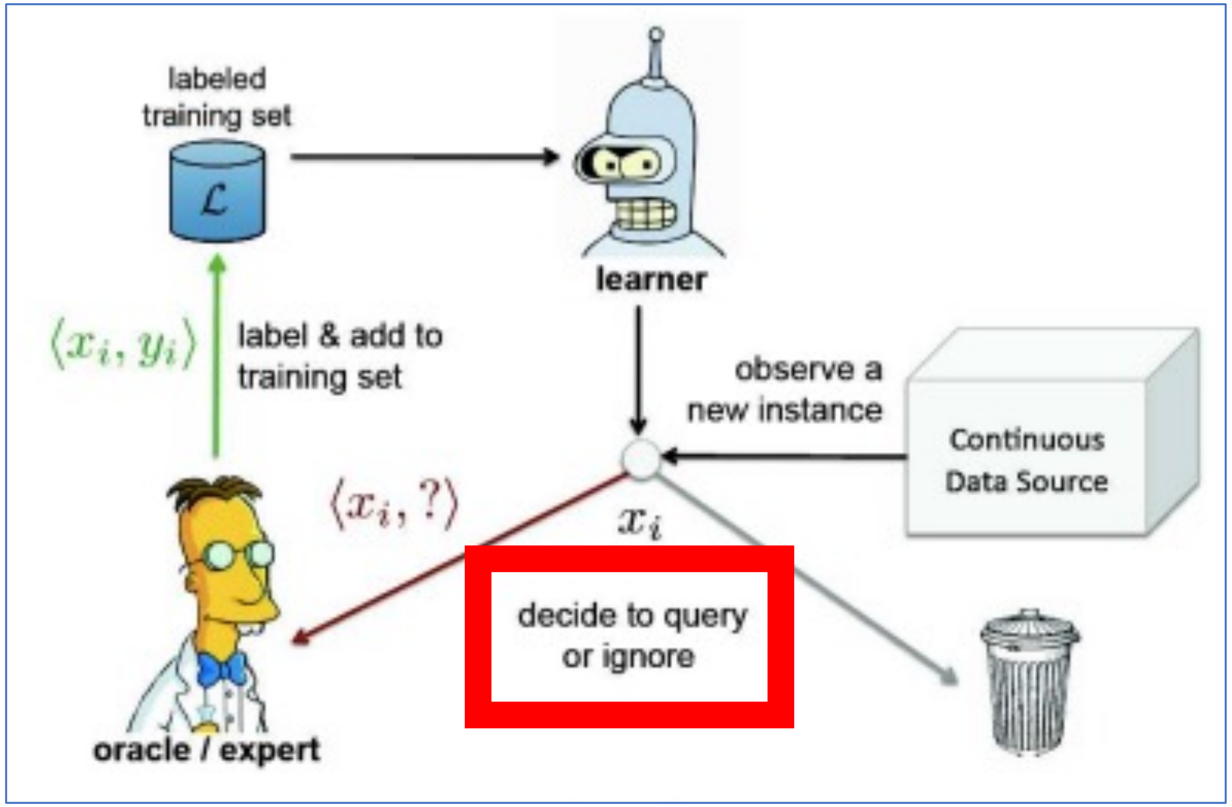
How to teach machines with minimal human supervision?

e.g., limited access to
(expert) annotators

e.g., limited funding

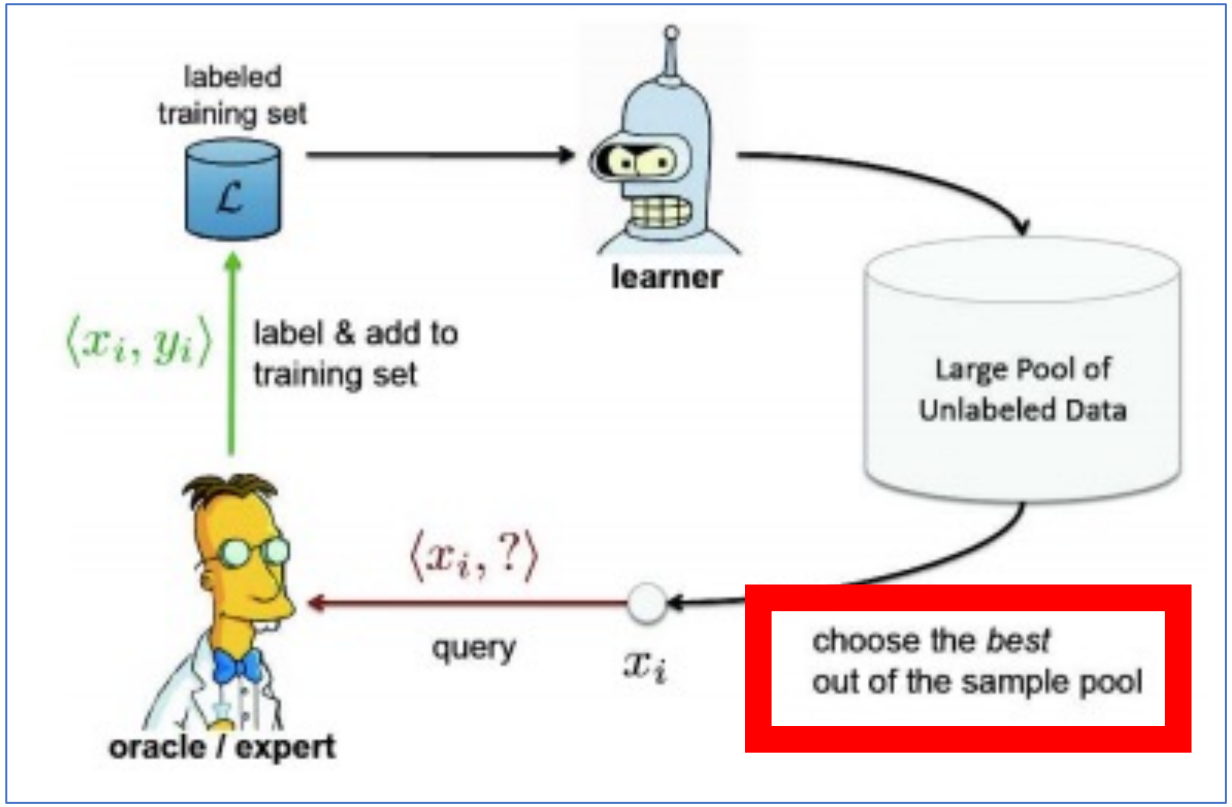
Idea: Choose Most Informative Data to Label

Stream-Based



Consider one example at a time

Pool-Based



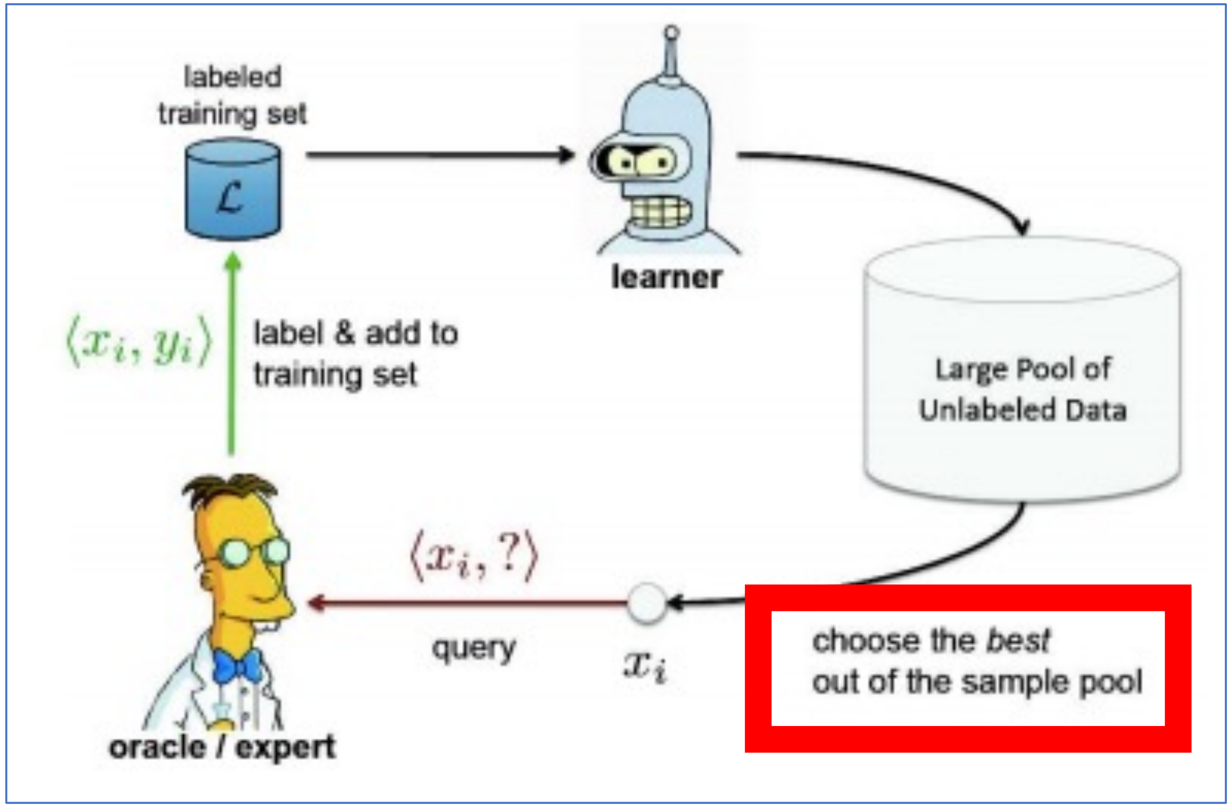
Consider many examples at a time

Image Credit: https://www.cs.utah.edu/~piyush/teaching/10-11-slides.pdf

Active Learning for Neural Networks: Status Quo

Iteratively add more labelled training examples after *n* epochs; different from curriculum learning because labels need to be collected for the added data

Pool-Based



Consider many examples at a time

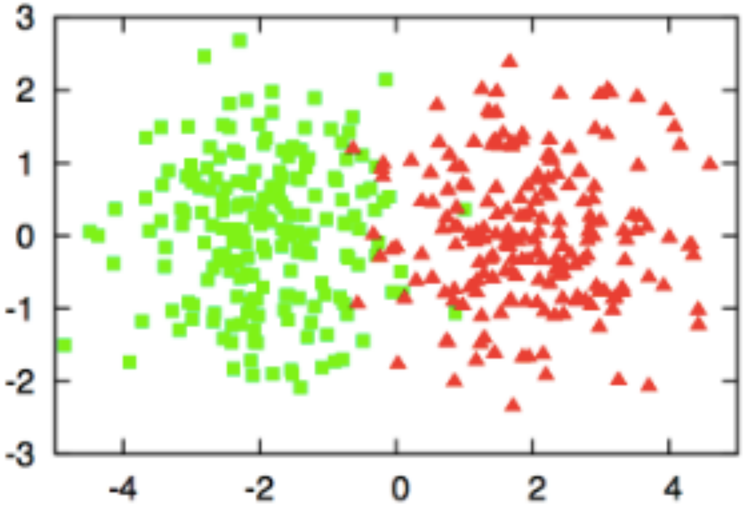
Image Credit: https://www.cs.utah.edu/~piyush/teaching/10-11-slides.pdf

What approach might be effective in identifying the most informative data to label?

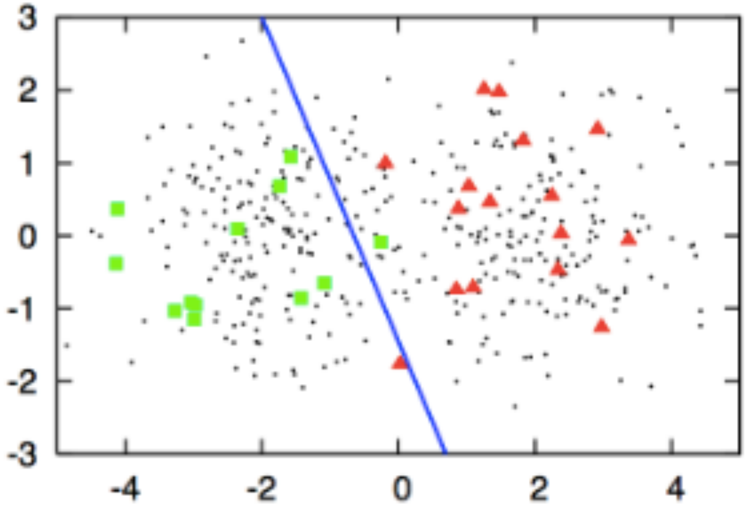
Common Approach: Uncertainty Sampling

Query instance(s) the classifier is most uncertain about.

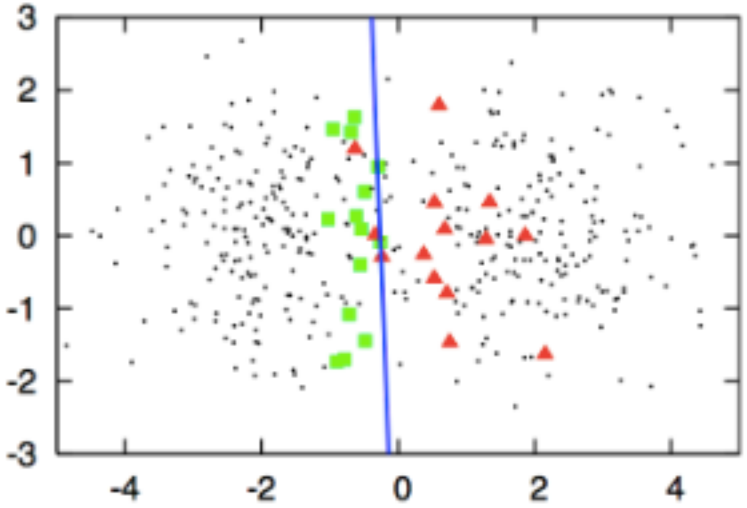
True Representation (Assume Labels Are Not Known)



Passive Learner (Random Selection)



Active Learner (Uncertainty Sampling)



http://burrsettles.com/pub/settles.activelearning.pdf

e.g., Uncertainty Estimation for Neural Networks Using Robustness Testing

Use model's predictions on random augmentations of the input to measure consistency/uncertainty; e.g.,

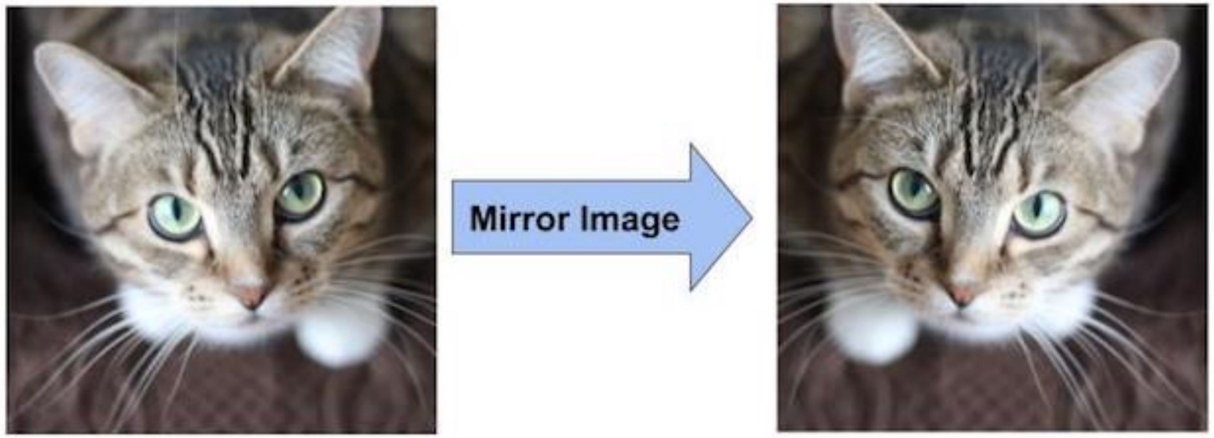
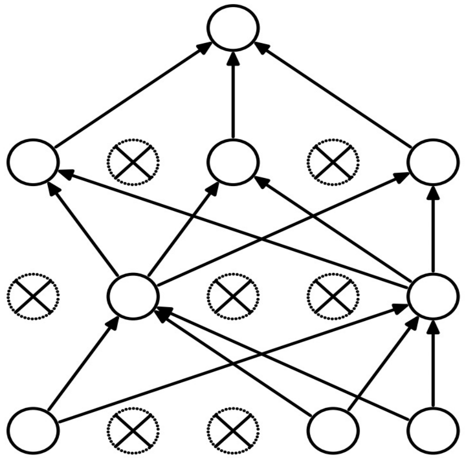


Figure Source: https://learnopencv.com/understanding-alexnet/

Elezi et al. Not all labels are equal: rationalizing the labeling costs for training object detection. CVPR 2022

e.g., Uncertainty Estimation for Neural Networks Using Ensembles (Two Approaches)

1. Dropout with different masks at inference time



(b) After applying dropout.

2. Multiple neural networks

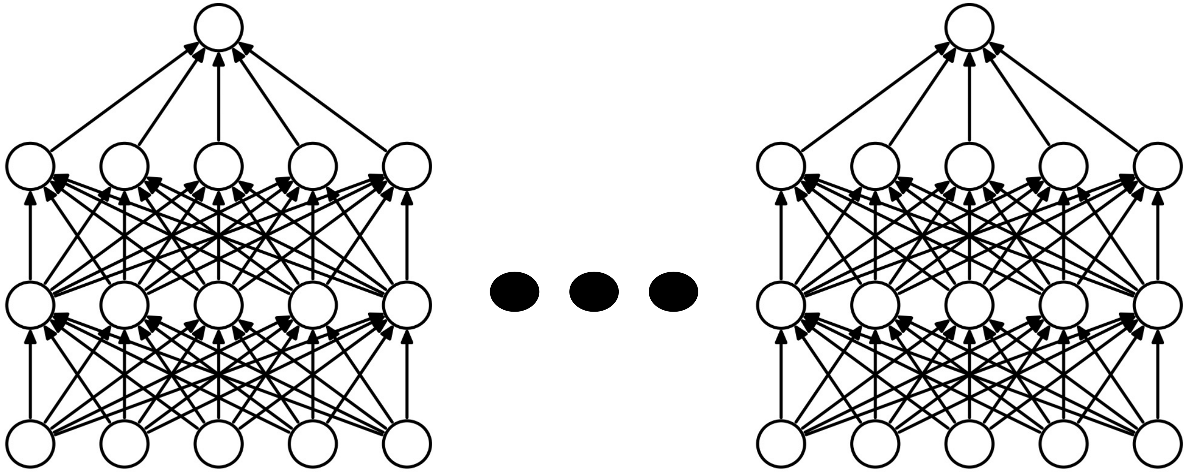


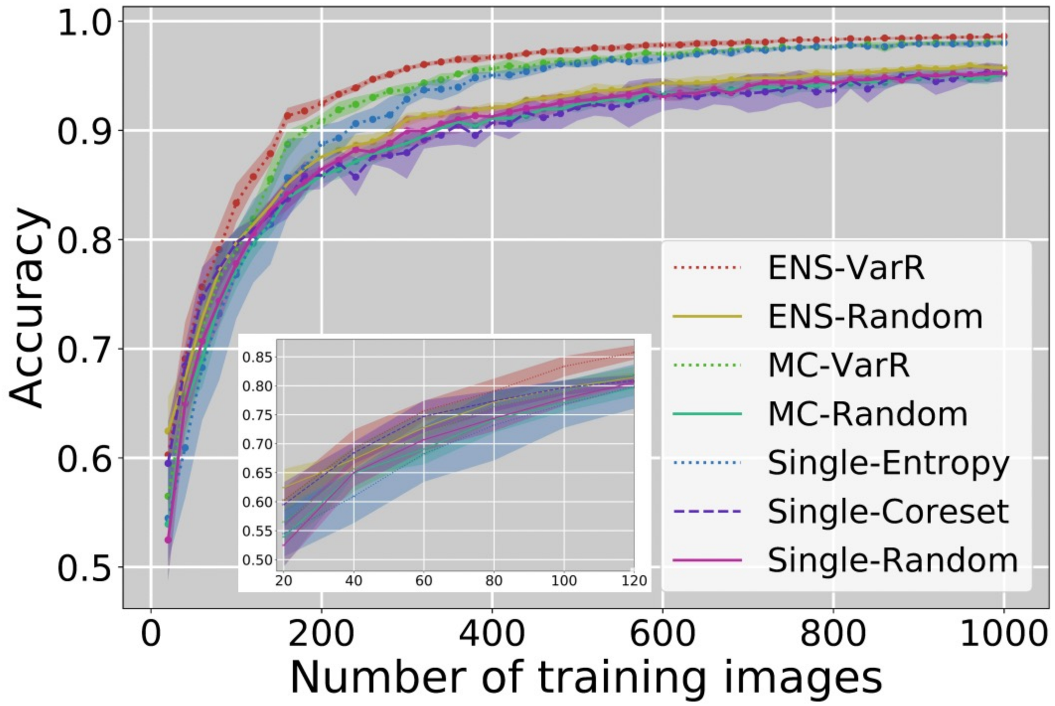
Figure Source: Srivastava et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research. 2014

Predicted softmax probabilities used to estimate uncertainty (e.g., entropy across softmax values), with average taken across all ensemble's softmax distributions

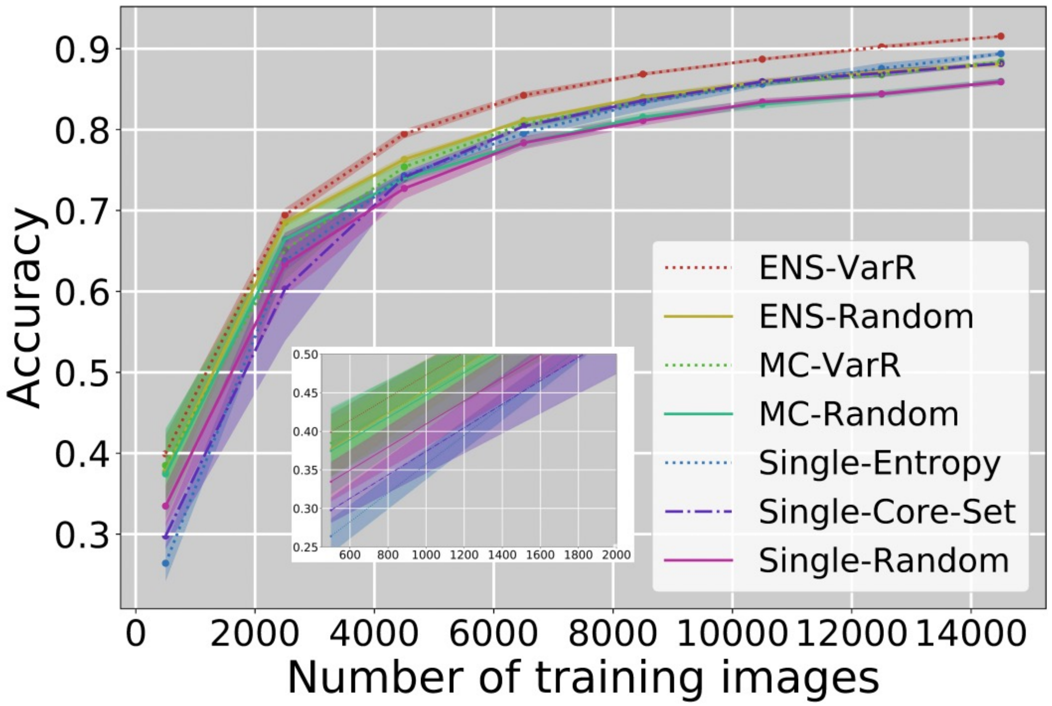
Beluch et al. The power of ensembles for active learning in image classification. CVPR 2018

e.g., Uncertainty Estimation for Neural Networks Using Ensembles (Two Approaches)

Active learning methods lead to faster learning and reduced human annotation effort than passive (random) learning for two image classification datasets



(a) MNIST on S-CNN

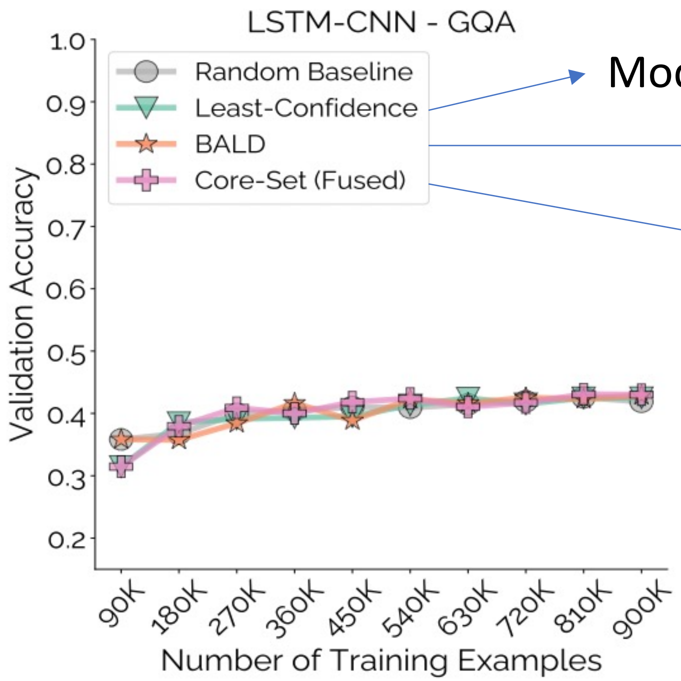


(b) CIFAR-10 on DenseNet

Beluch et al. The power of ensembles for active learning in image classification. CVPR 2018

Common AL Techniques Have Mixed Results

- **Successes**: image classification, object detection
- **Failure**: VQA (e.g., AL methods label 10% of overall pool per iteration; initial model trained on 10% of pool)



Model's confidence in prediction

Based on ensemble from dropout

Examples capturing diversity in data pool

Karamcheti et al. Mind your outliers! Investigating the negative impact of outliers on active learning for visual question answering. Association for Computational Linguistics (ACL) 2021

Common AL Techniques Have Mixed Results

Why might AL methods perform comparable or worse to random selection?

- Challenging examples to learn are sampled; e.g.,

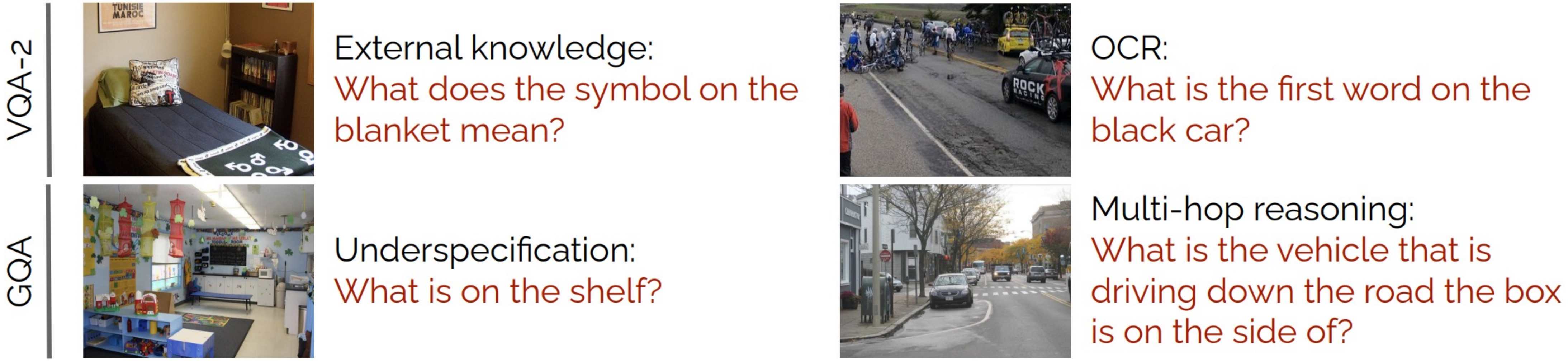
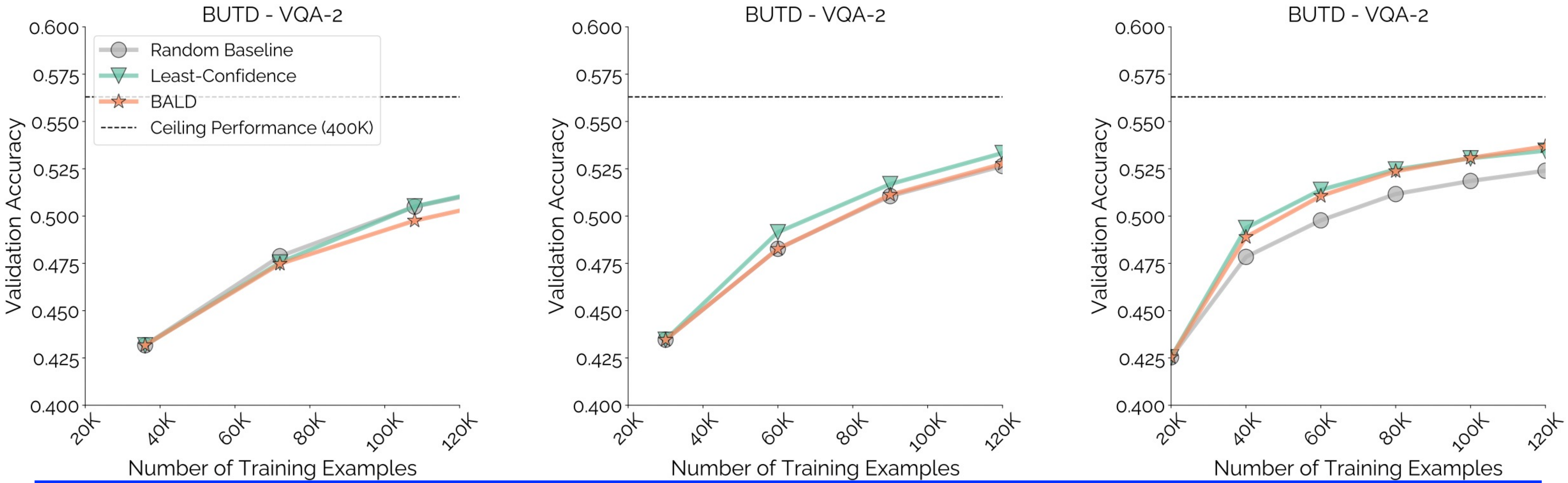


Figure 7: Example groups of collective outliers in the VQA-2 and GQA datasets.

Karamcheti et al. Mind your outliers! Investigating the negative impact of outliers on active learning for visual question answering. Association for Computational Linguistics (ACL) 2021

Idea: Remove “Unlearnable” Data from Pool

Performance compared to random selection improves for AL approaches when removing “challenging” examples from data pool



(a) 10% of Dataset Removed (b) 25% of Dataset Removed (c) 50% of Dataset Removed

Karamcheti et al. Mind your outliers! Investigating the negative impact of outliers on active learning for visual question answering. Association for Computational Linguistics (ACL) 2021

Recent Works: ICCV 2023 Papers; e.g.,

Heterogeneous Diversity Driven Active Learning for Multi-Object Tracking

ng[1,†]

HAL3D: Hierarchical Active Learning for Fine-Grained 3D Part Labeling

ALWOD: Active Learning for Weakly-Supervised Object Detection

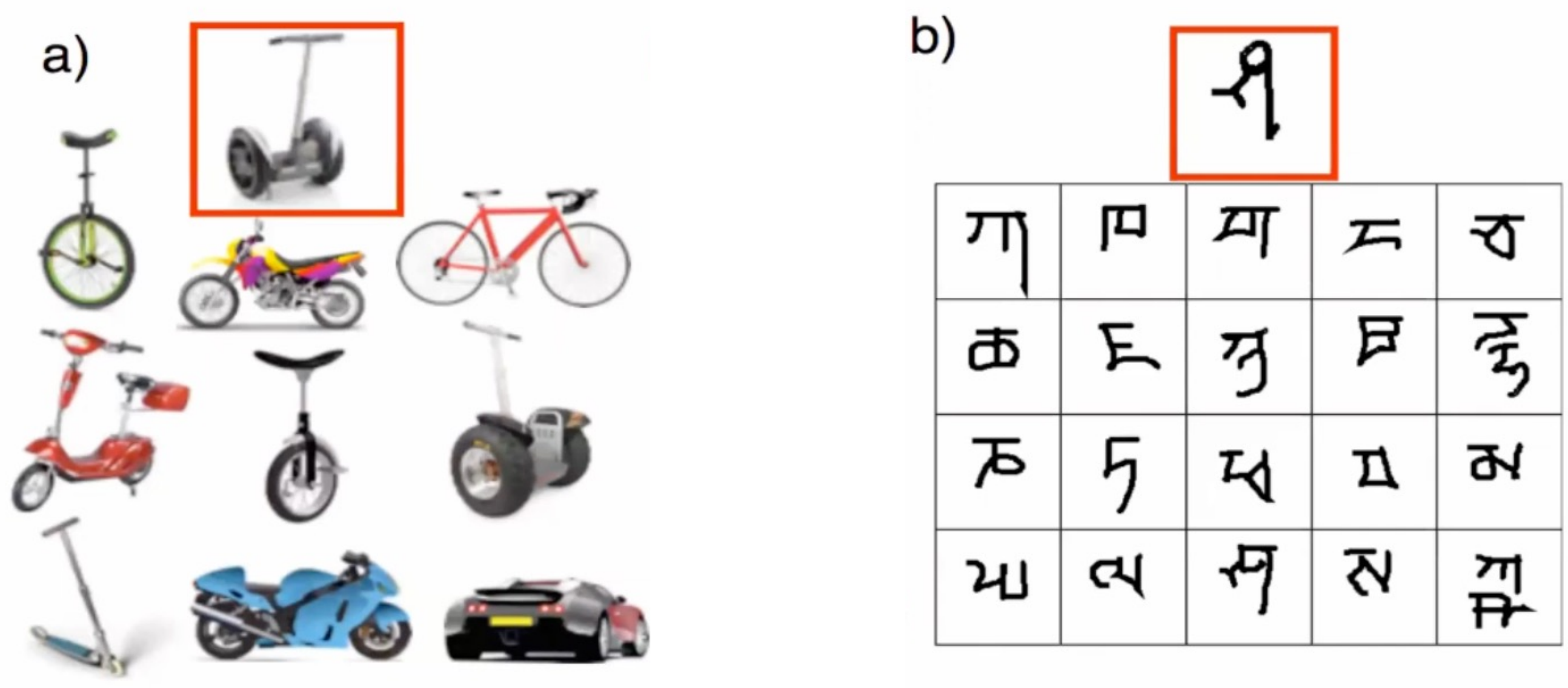
Yuting Wang[1], Velibor Ilic[2], Jiatong Li[1], Branislav Kisačanin[3,2], and Vladimir Pavlovic[1]

[1]Rutgers University, NJ, USA
[2]The Institute for Artificial Intelligence Research and Development of Serbia, Novi Sad, Serbia
[3]Nvidia Corporation, TX, USA
yw632@rutgers.edu, velibor.ilic@ivi.ac.rs, jiatong.li@rutgers.edu,
b.kisacanin@ieee.org, vladimir@cs.rutgers.edu

Efficient Learning: Today’s Topics

- Motivation
- Curriculum Learning
- Active Learning
- Few-shot Learning
- Faculty Course Questionnaire (FCQ)

Intuition: Generalize Current Knowledge

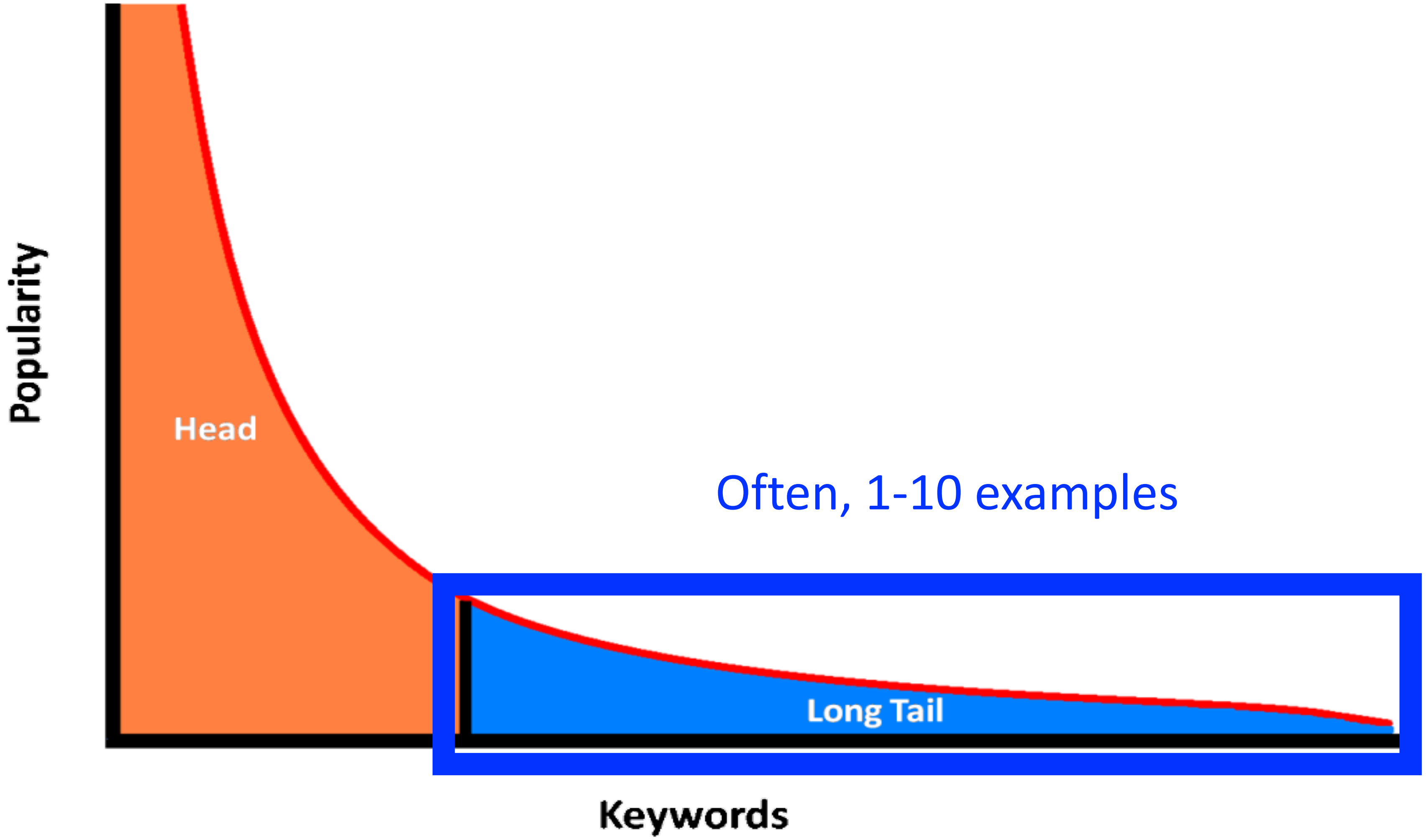


Lake et al, 2013, 2015

Given one example per category, identify the category of the query

https://www.youtube.com/watch?v=9j4iH9TPTd8

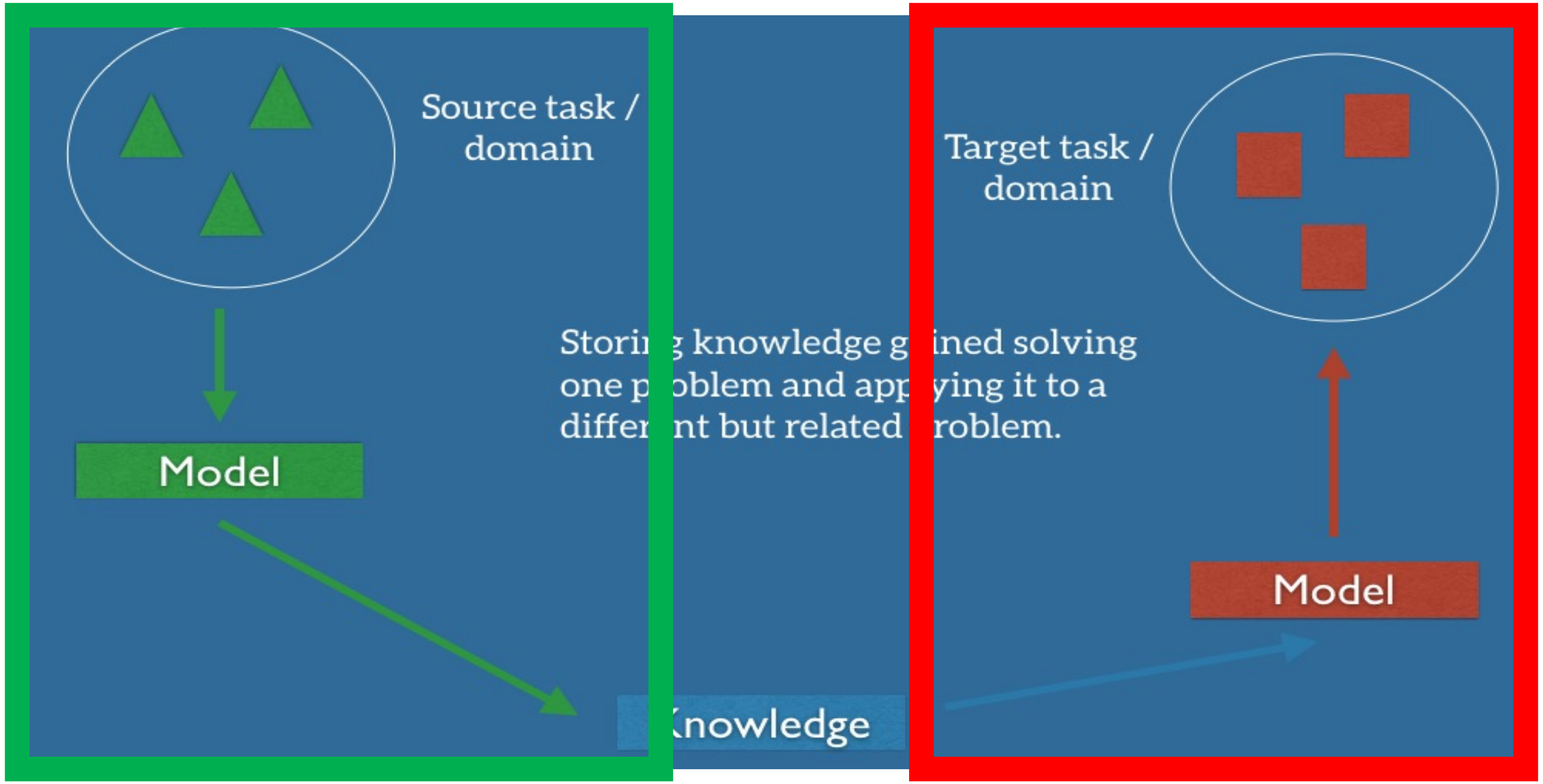
Problem Set-up: Learn from Few Examples



https://daredevilmusicproduction.com/long-tail/

Problem Set-up: Learn from Few Examples

Sufficient Labeled Data



Insufficient Labeled Data

- Few shot learning: evaluate only for categories with few examples
- Generalized few shot learning: evaluate on all categories

https://ruder.io/transfer-learning/

What are applications for which we might have limited examples?

- medical
- outer space
- natural disasters

Popular Approaches

- Design-time approach: fine-tuning
- Run-time approach: meta learning

Popular Approaches

- Design-time approach: fine-tuning
- Run-time approach: meta learning

Recall Fine-Tuning

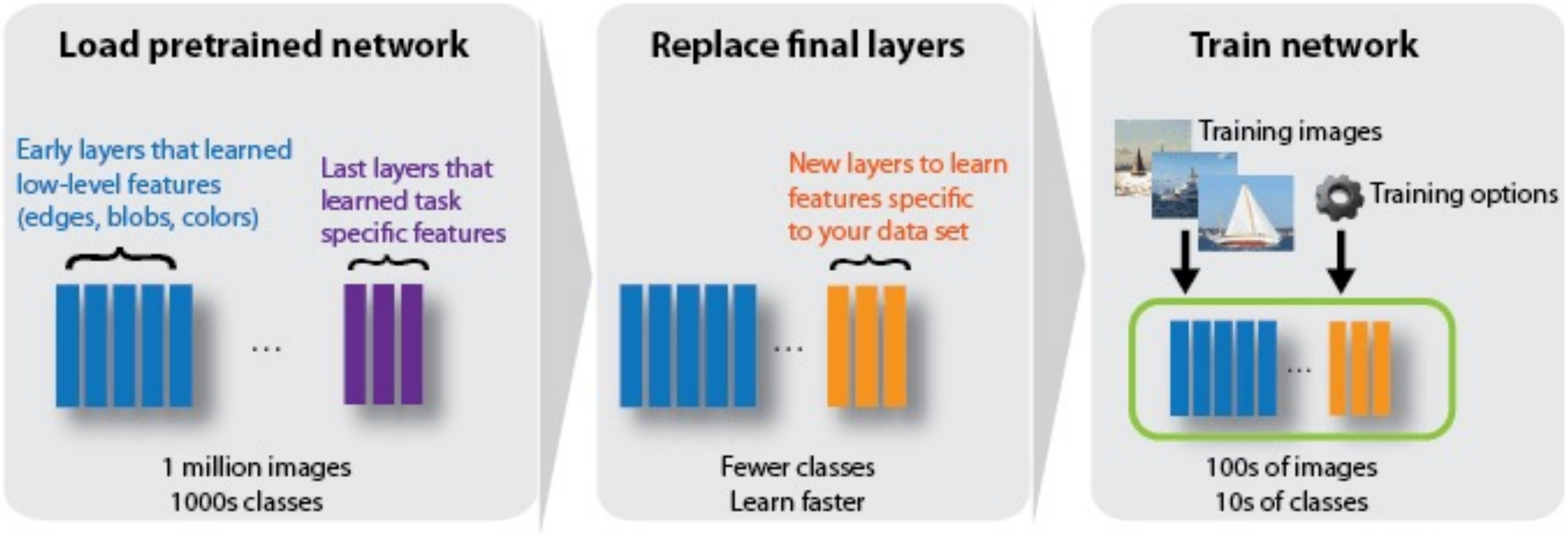
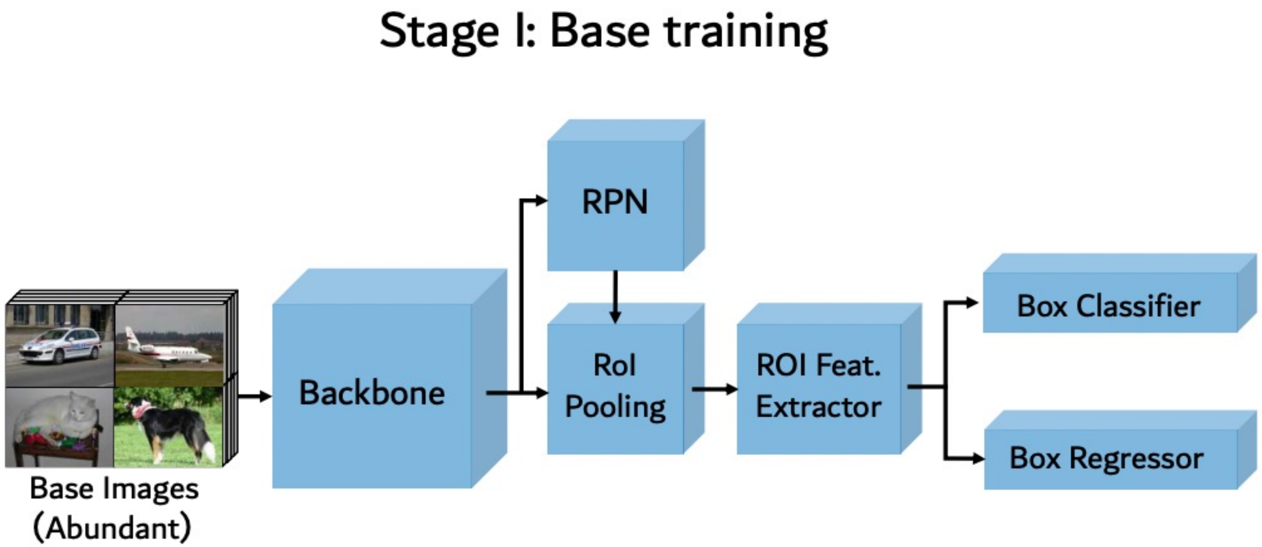
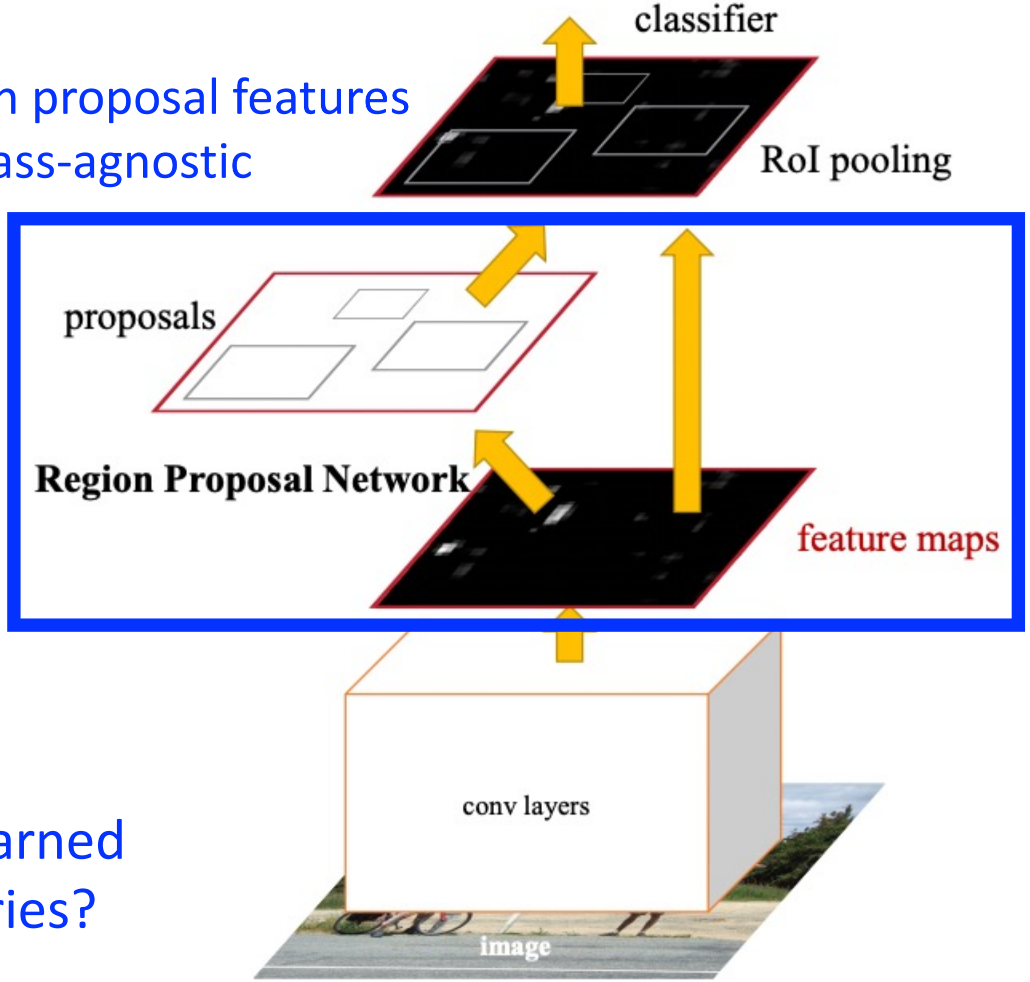


Image Source: https://www.mathworks.com/help/deeplearning/ug/transfer-learning-using-alexnet.html

e.g., Fine-Tuning for Object Detection

Region proposal features are class-agnostic

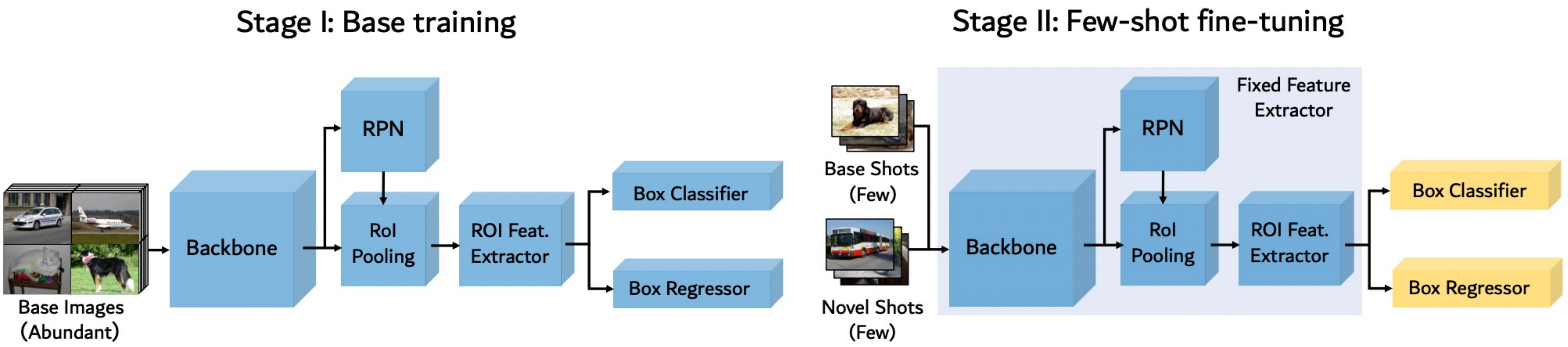




Faster R-CNN architecture: Why would we anticipate learned features would generalize well to locating novel categories?

Ren Shaoqing Ren et al. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.” Neurips 2015.
Wang et al. Frustratingly simple few-shot object detection. arXiv 2020.

e.g., Fine-Tuning for Object Detection



K shots from both base and novel categories used for training

Why include shots from both base and novel categories?

Wang et al. Frustratingly simple few-shot object detection. arXiv 2020.

e.g., Fine-Tuning for Object Detection

Tested with cross validation on 3 splits from VOC

mAP scores for training with 1, 2, 3, 5, and 10 examples (shots) per category

Method / Shot	Backbone	Novel Set 1					Novel Set 2					Novel Set 3				
		1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
YOLO-joint (Kang et al., 2019)	YOLOv2	0.0	0.0	1.8	1.8	1.8	0.0	0.1	0.0	1.8	0.0	1.8	1.8	1.8	3.6	3.9
YOLO-ft (Kang et al., 2019)	YOLOv2	3.2	6.5	6.4	7.5	12.3	8.2	3.8	3.5	3.5	7.8	8.1	7.4	7.6	9.5	10.5
YOLO-ft-full (Kang et al., 2019)	YOLOv2	6.6	10.7	12.5	24.8	38.6	12.5	4.2	11.6	16.1	33.9	13.0	15.9	15.0	32.2	38.4
FSRW (Kang et al., 2019)	YOLOv2	14.8	15.5	26.7	33.9	47.2	15.7	15.3	22.7	30.1	40.5	21.3	25.6	28.4	42.8	45.9
MetaDet (Wang et al., 2019b)	YOLOv2	17.1	19.1	28.9	35.0	48.8	18.2	20.6	25.9	30.6	41.5	20.1	22.3	27.9	41.9	42.9
FRCN+joint (Wang et al., 2019b)	FRCN w/VGG16	0.3	0.0	1.2	0.9	1.7	0.0	0.0	1.1	1.9	1.7	0.2	0.5	1.2	1.9	2.8
FRCN+joint-ft (Wang et al., 2019b)	FRCN w/VGG16	9.1	10.9	13.7	25.0	39.5	10.9	13.2	17.6	19.5	36.5	15.0	15.1	18.3	33.1	35.9
MetaDet (Wang et al., 2019b)	FRCN w/VGG16	18.9	20.6	30.2	36.8	49.6	21.8	23.1	27.8	31.7	43.0	20.6	23.9	29.4	43.9	44.1
FRCN+joint (Yan et al., 2019)	FRCN w/R-101	2.7	3.1	4.3	11.8	29.0	1.9	2.6	8.1	9.9	12.6	5.2	7.5	6.4	6.4	6.4
FRCN+ft (Yan et al., 2019)	FRCN w/R-101	11.9	16.4	29.0	36.9	36.9	5.9	8.5	23.4	29.1	28.8	5.0	9.6	18.1	30.8	43.4
FRCN+ft-full (Yan et al., 2019)	FRCN w/R-101	13.8	19.6	32.8	41.5	45.6	7.9	15.3	26.2	31.6	39.1	9.8	11.3	19.1	35.0	45.1
Meta R-CNN (Yan et al., 2019)	FRCN w/R-101	19.9	25.5	35.0	45.7	51.5	10.4	19.4	29.6	34.8	**45.4**	14.3	18.2	27.5	41.2	48.1
FRCN+ft-full (Our Impl.)	FRCN w/R-101	15.2	20.3	29.0	40.1	45.5	13.4	20.6	28.6	32.4	38.8	19.6	20.8	28.7	42.2	42.1
TFA w/ fc (Ours)	FRCN w/R-101	36.8	29.1	43.6	**55.7**	**57.0**	18.2	**29.0**	33.4	**35.5**	39.0	27.7	33.6	42.5	48.7	**50.2**
TFA w/ cos (Ours)	FRCN w/R-101	**39.8**	**36.1**	**44.7**	**55.7**	56.0	**23.5**	26.9	**34.1**	35.1	39.1	**30.8**	**34.8**	**42.8**	**49.5**	49.8

Consistently outperforms baselines by 2-20 points on novel categories

Wang et al. Frustratingly simple few-shot object detection. arXiv 2020.

e.g., Fine-Tuning for Object Detection

Tested with cross validation on 3 splits from VOC

mAP scores for training with 1, 2, 3, 5, and 10 examples (shots) per category

Method / Shot	Backbone	Novel Set 1					Novel Set 2					Novel Set 3				
		1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
YOLO-joint (Kang et al., 2019)	YOLOv2	0.0	0.0	1.8	1.8	1.8	0.0	0.1	0.0	1.8	0.0	1.8	1.8	1.8	3.6	3.9
YOLO-ft (Kang et al., 2019)		3.2	6.5	6.4	7.5	12.3	8.2	3.8	3.5	3.5	7.8	8.1	7.4	7.6	9.5	10.5
YOLO-ft-full (Kang et al., 2019)		6.6	10.7	12.5	24.8	38.6	12.5	4.2	11.6	16.1	33.9	13.0	15.9	15.0	32.2	38.4
FSRW (Kang et al., 2019)		14.8	15.5	26.7	33.9	47.2	15.7	15.3	22.7	30.1	40.5	21.3	25.6	28.4	42.8	45.9
MetaDet (Wang et al., 2019b)		17.1	19.1	28.9	35.0	48.8	18.2	20.6	25.9	30.6	41.5	20.1	22.3	27.9	41.9	42.9
FRCN+joint (Wang et al., 2019b)	FRCN w/VGG16	0.3	0.0	1.2	0.9	1.7	0.0	0.0	1.1	1.9	1.7	0.2	0.5	1.2	1.9	2.8
FRCN+joint-ft (Wang et al., 2019b)		9.1	10.9	13.7	25.0	39.5	10.9	13.2	17.6	19.5	36.5	15.0	15.1	18.3	33.1	35.9
MetaDet (Wang et al., 2019b)		18.9	20.6	30.2	36.8	49.6	21.8	23.1	27.8	31.7	43.0	20.6	23.9	29.4	43.9	44.1
FRCN+joint (Yan et al., 2019)	FRCN w/R-101	2.7	3.1	4.3	11.8	29.0	1.9	2.6	8.1	9.9	12.6	5.2	7.5	6.4	6.4	6.4
FRCN+ft (Yan et al., 2019)		11.9	16.4	29.0	36.9	36.9	5.9	8.5	23.4	29.1	28.8	5.0	9.6	18.1	30.8	43.4
FRCN+ft-full (Yan et al., 2019)		13.8	19.6	32.8	41.5	45.6	7.9	15.3	26.2	31.6	39.1	9.8	11.3	19.1	35.0	45.1
Meta R-CNN (Yan et al., 2019)		19.9	25.5	35.0	45.7	51.5	10.4	19.4	29.6	34.8	**45.4**	14.3	18.2	27.5	41.2	48.1
FRCN+ft-full (Our Impl.)	FRCN w/R-101	15.2	20.3	29.0	40.1	45.5	13.4	20.6	28.6	32.4	38.8	19.6	20.8	28.7	42.2	42.1
TFA w/ fc (Ours)		36.8	29.1	43.6	**55.7**	**57.0**	18.2	**29.0**	33.4	**35.5**	39.0	27.7	33.6	42.5	48.7	**50.2**
TFA w/ cos (Ours)		**39.8**	**36.1**	**44.7**	**55.7**	56.0	**23.5**	26.9	**34.1**	35.1	39.1	**30.8**	**34.8**	**42.8**	**49.5**	49.8

Similar performance boosts also observed on two more datasets (COCO and LVIS)

Wang et al. Frustratingly simple few-shot object detection. arXiv 2020.

Fine-Tuning

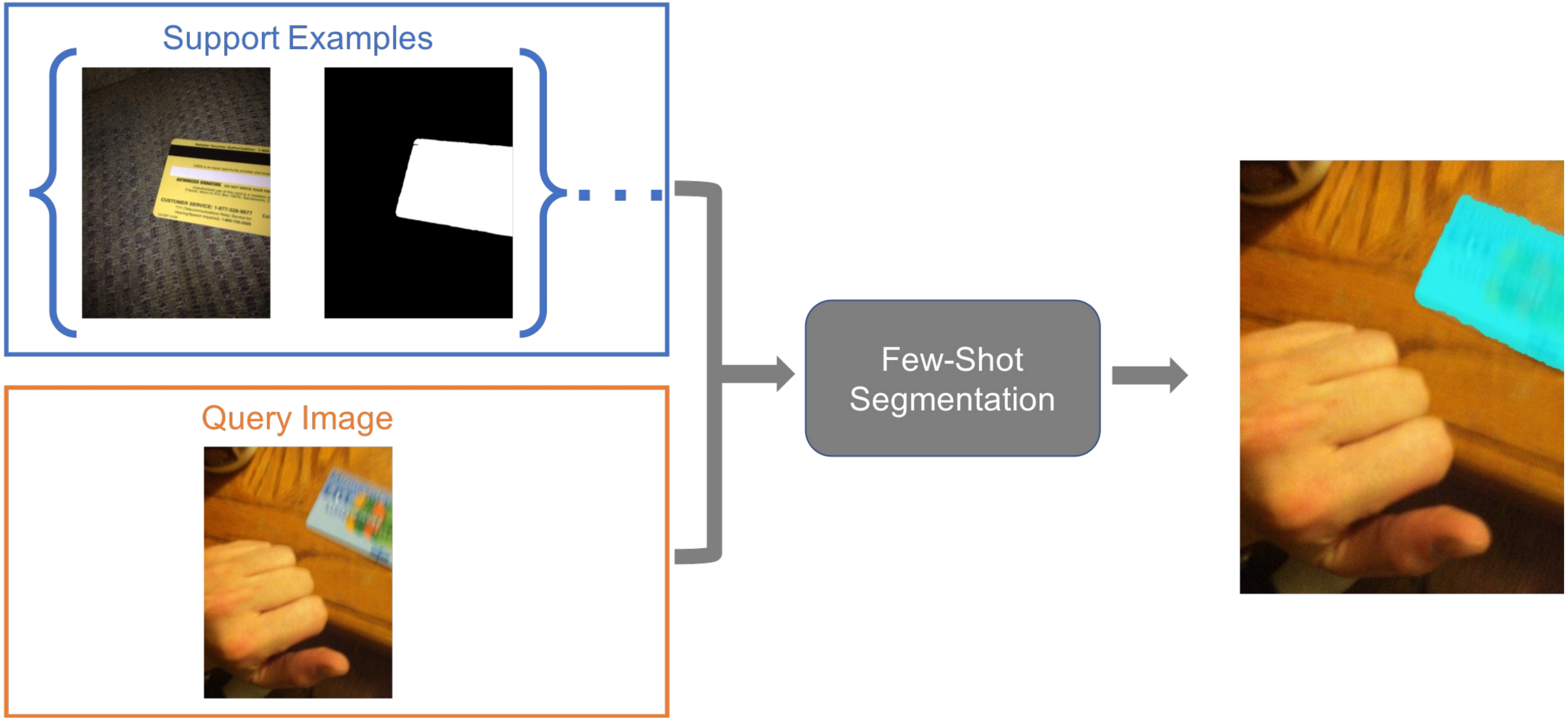
What are limitations of this approach for real-world applications?

- Must retrain algorithm to add new categories

Popular Approaches

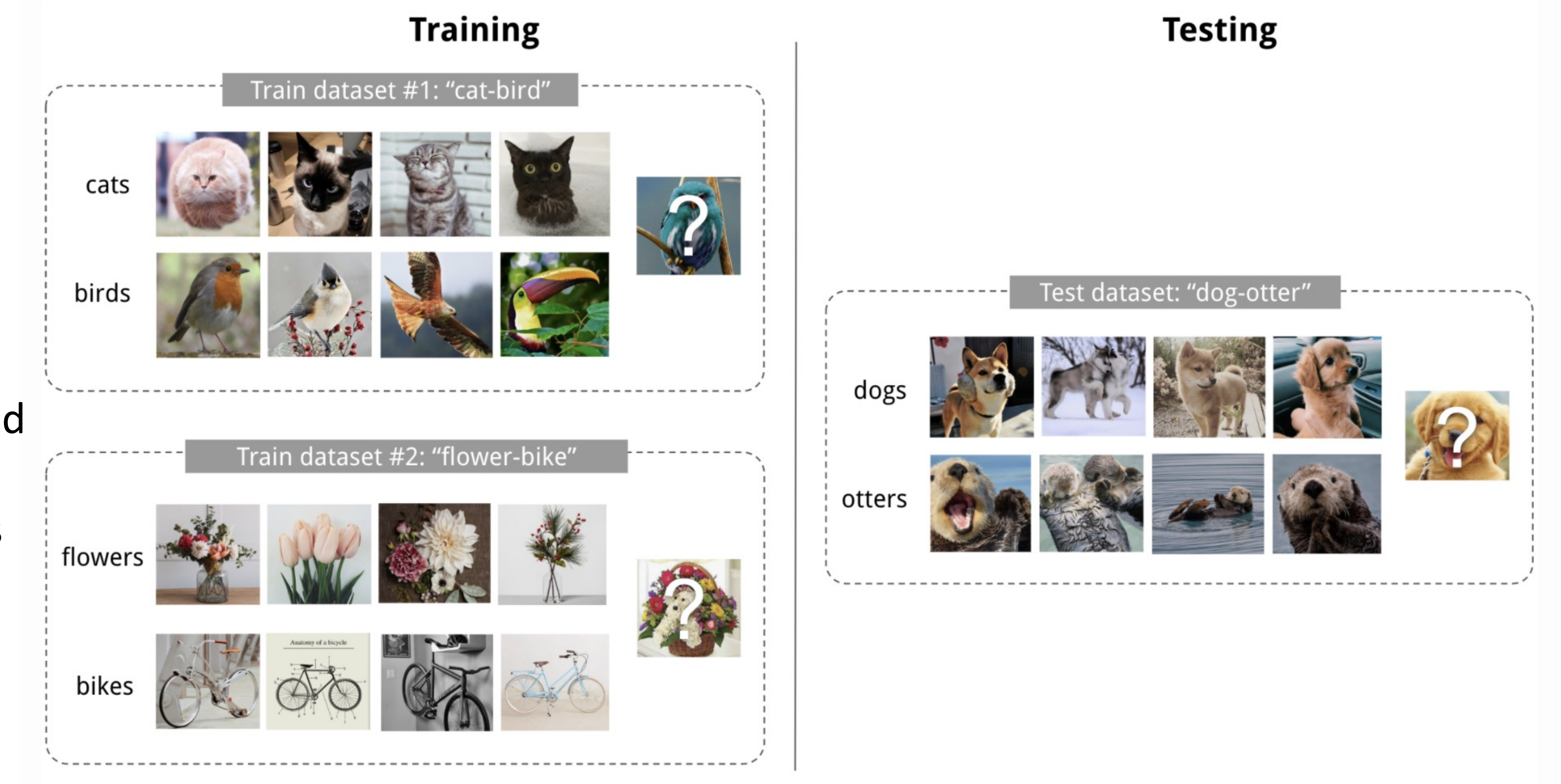
- Design-time approach: fine-tuning
- Run-time approach: meta learning

Meta Learner: Update Model with Support Set



Implementation: Trained Model Updates Itself to Generalize to Support Set Categories

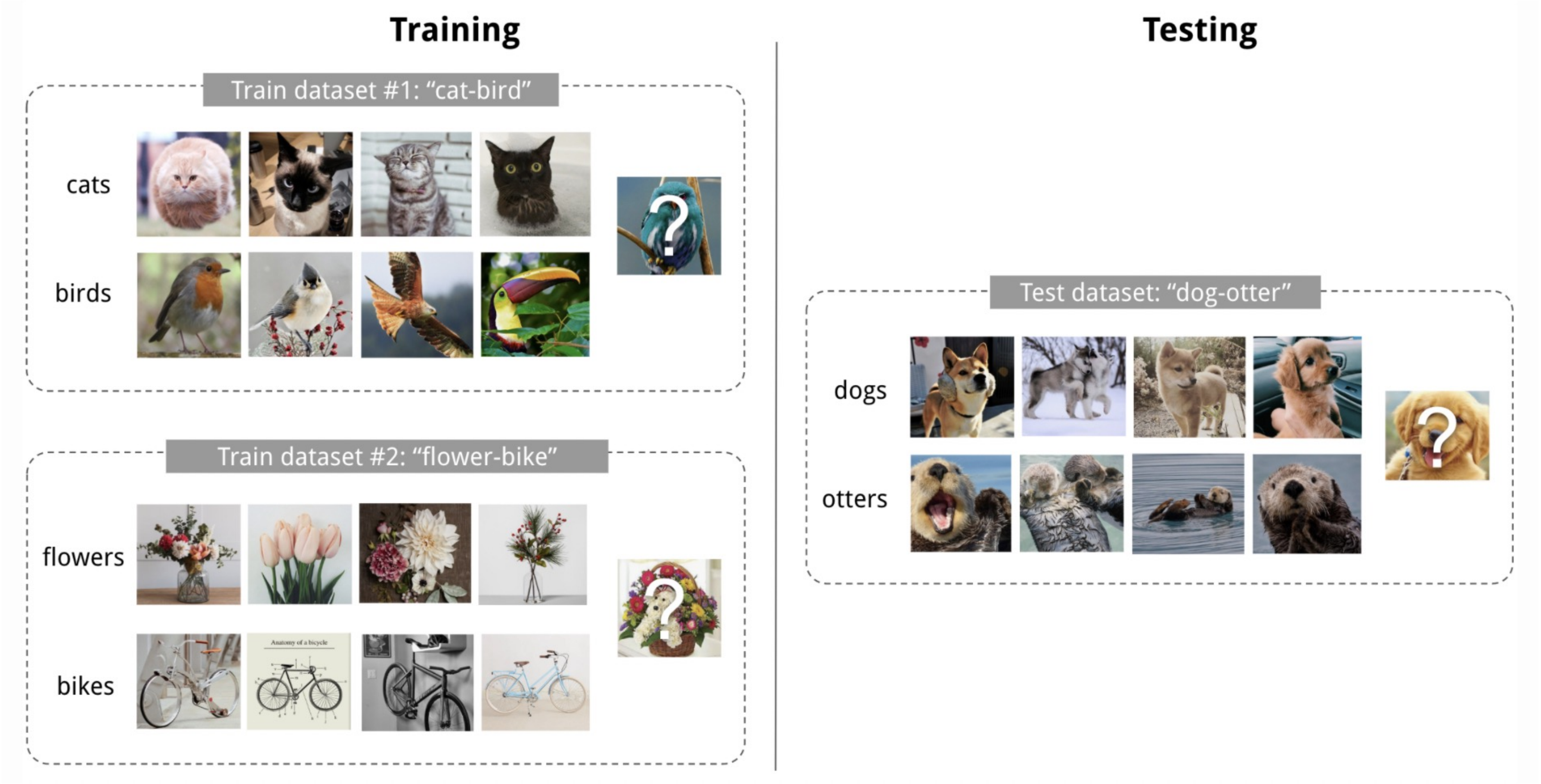
Goal: learn features during training that are class-agnostic and so can generalize to novel test categories



https://lilianweng.github.io/lil-log/2018/11/30/meta-learning.html

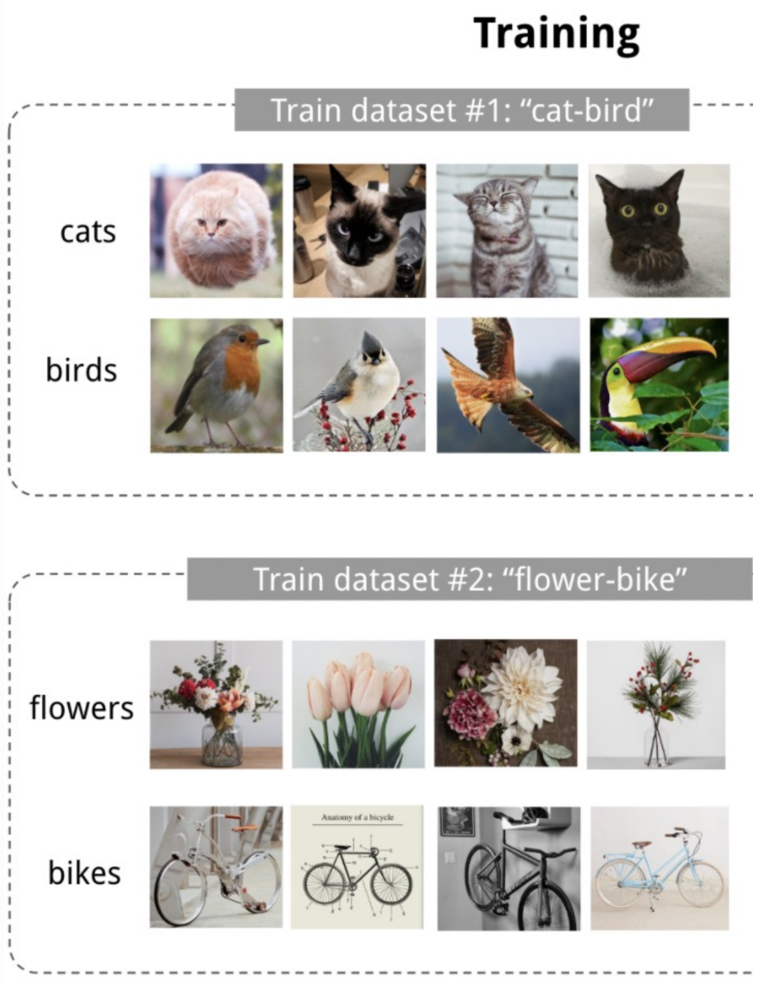
Implementation: Trained Model Updates Itself to Generalize to Support Set Categories

How many shots are observed at testing?



https://lilianweng.github.io/lil-log/2018/11/30/meta-learning.html

Implementation: Trained Model Updates Itself to Generalize to Support Set Categories



How many "shots" should be observed at each training round?

- 4 (must match test time)

https://lilianweng.github.io/lil-log/2018/11/30/meta-learning.html

Implementation: Trained Model Updates Itself to Generalize to Support Set Categories

Given support categories, detect which one the "query" matches

Recall support categories are never observed during training!

https://lilianweng.github.io/lil-log/2018/11/30/meta-learning.html

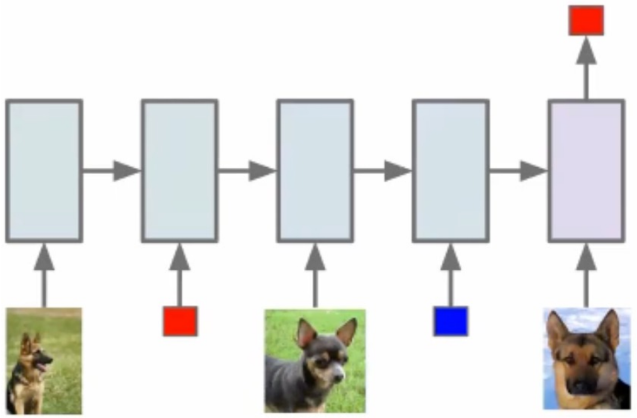
Implementation: Trained Model Updates Itself to Generalize to Support Set Categories

How to train a model to do this?

https://lilianweng.github.io/lil-log/2018/11/30/meta-learning.html

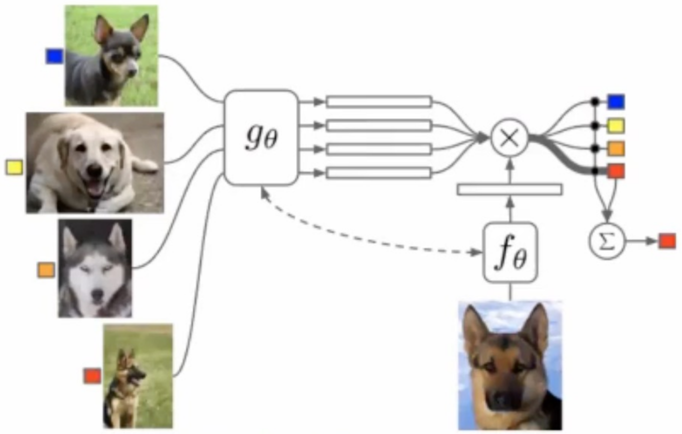
Implementation: Trained Model Updates Itself to Generalize to Support Set Categories

Model Based



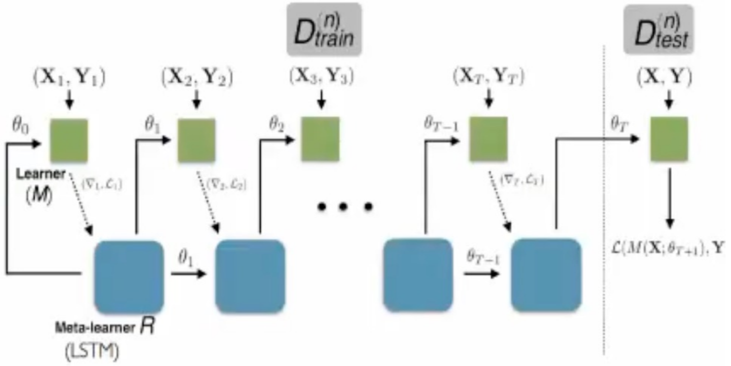
- Santoro et al. '16
- Duan et al. '17
- Wang et al. '17
- Munkhdalai & Yu '17
- Mishra et al. '17
- ...

Metric Based



- Koch '15
- Vinyals et al. '16
- Snell et al. '17
- Shyam et al. '17
- Sung et al. '17
- ...

Optimization Based



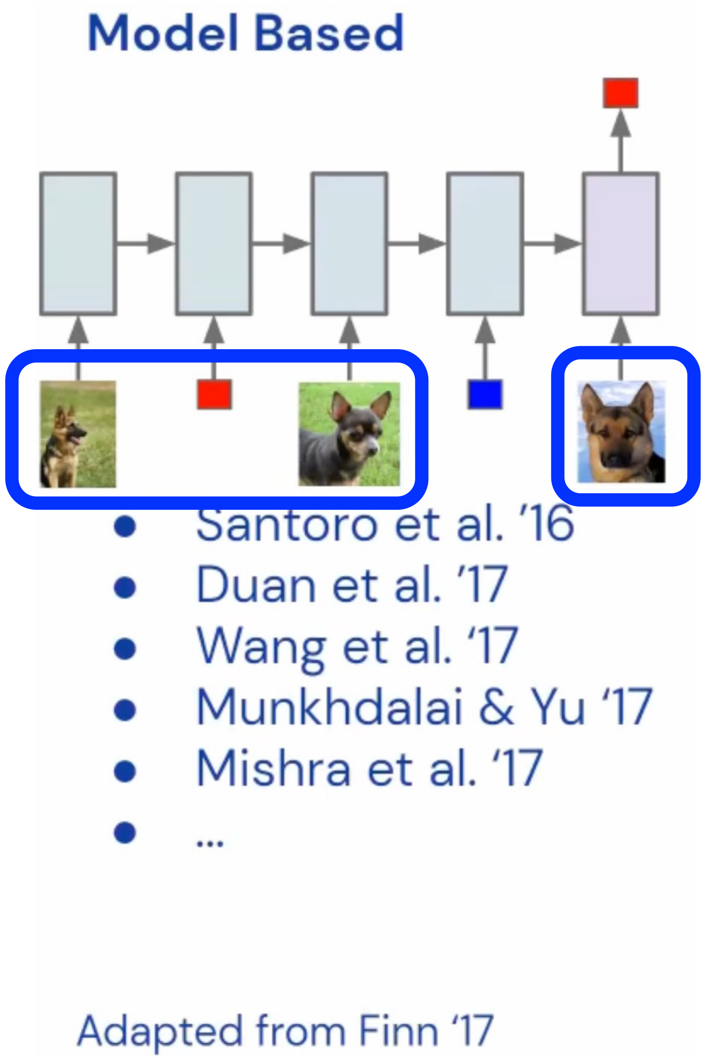
- Schmidhuber '87, '92
- Bengio et al. '90, '92
- Hochreiter et al. '01
- Li & Malik '16
- Andrychowicz et al. '16
- Ravi & Larochelle '17
- Finn et al. '17
- ...

Adapted from Finn '17

https://www.youtube.com/watch?v=9j4iH9TPTd8

Implementation: Trained Model Updates Itself to Generalize to Support Set Categories

Support: data + label



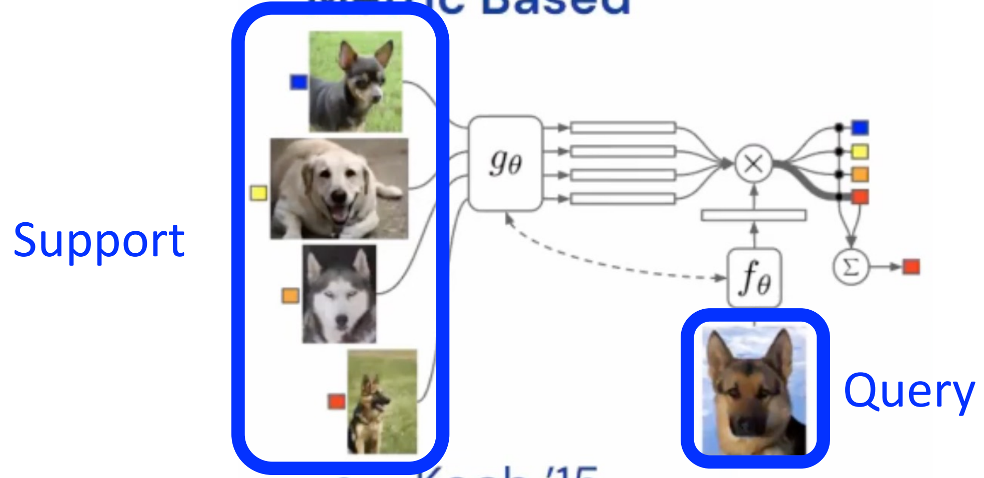
Query

e.g., learn set-invariant neural networks, such as those that rely on attention, to locate similarity

https://www.youtube.com/watch?v=9j4iH9TPTd8

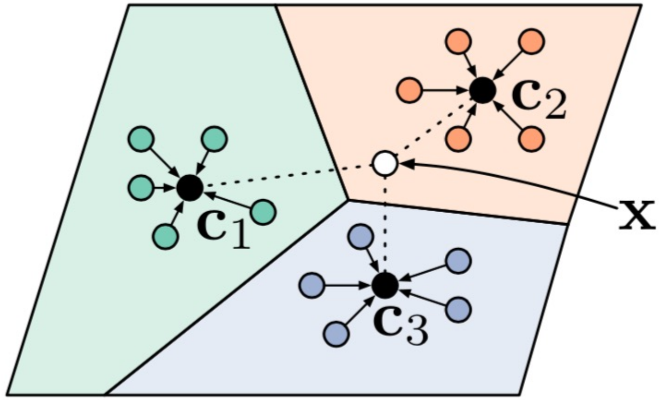
Implementation: Trained Model Updates Itself to Generalize to Support Set Categories

Metric Based



- Koch '15
- Vinyals et al. '16
- Snell et al. '17
- Shyam et al. '17
- Sung et al. '17
- ...

Compare query to each support category; e.g., establish a "prototype" for each support set

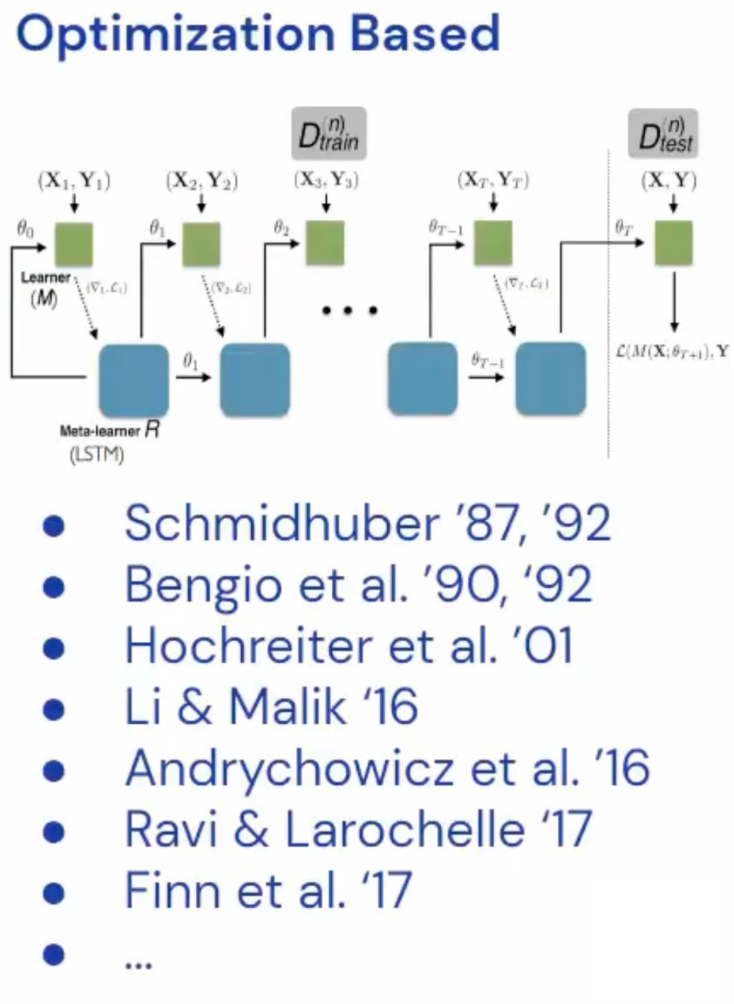


https://lilianweng.github.io/posts/2018-11-30-meta-learning/

https://www.youtube.com/watch?v=9j4iH9TPTd8

Implementation: Trained Model Updates Itself to Generalize to Support Set Categories

Function to optimize is conditioned on the support set; e.g., tweak “forget” gate of LSTM



https://www.youtube.com/watch?v=9j4iH9TPTd8

Meta Learner: Update Model with Support Set

What are limitations of this approach for real-world applications?

- Requires large amount of memory to process the support set on top of the query set

Bronskill et al. Memory Efficient Meta-Learning with Large Images. Neurips 2021.

Popular Approaches

- Design-time approach: fine-tuning
- Run-time approach: meta learning

Efficient Learning: Today’s Topics

- Motivation
- Curriculum Learning
- Active Learning
- Few-shot Learning
- Faculty Course Questionnaire (FCQ)


The End