Efficient Learning

Danna Gurari
University of Colorado Boulder
Fall 2022

Review

- Last week:
  - Motivation
  - Key idea: knowledge distillation
  - Knowledge distillation for CNNs (vision problems)
  - Knowledge distillation for Transformers (language problems)

- Assignments (Canvas):
  - Final project presentations due Monday
    - Note: provide video URLs with YouTube or Vimeo

- Questions?

Today’s Topics

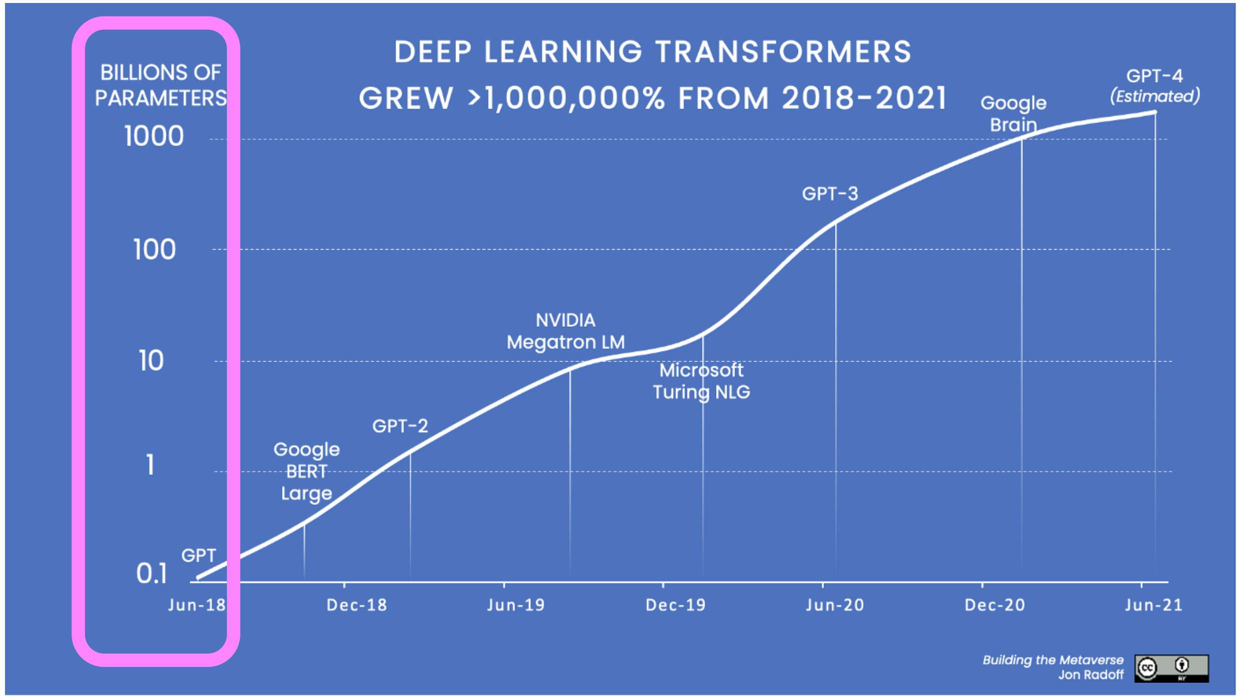
- Motivation
- Efficient learning: curriculum learning
- Efficient learning: active learning
- Efficient learning: other considerations
- Faculty course questionnaire

Today's Topics

- Motivation
- Efficient learning: curriculum learning
- Efficient learning: active learning
- Efficient learning: other considerations
- Faculty course questionnaire

Trend: Parameter-Heavy Models

Language – pretrained transformers



https://medium.com/building-the-metaverse/the-metaverse-and-artificial-intelligence-ai-577343895411

Vision – ImageNet classification

Architecture	Year	Top-1 Accuracy	# Parameters
DenseNet-169	2017	76.2%	14M
Inception-v3	2016	78.8%	24M
Inception-resnet-v2	2017	80.1%	56M
PolyNet	2017	81.3%	92M
SENet	2018	82.7%	146M
GPipe	2018	84.3%	557M
ResNeXt-101 32x48d	2019	85.4%	829M

Trend: Parameter-Heavy Models Are Often Predicated on Extensive Training

(Measured on Nvidia A100)

Models	#Params (M)	Training Time (GPU Hours)
ResNet-50	26	31
ResNet-101	45	44
BERT-Base	108	84
Turing-NLG 17B	17,000	TBA
GPT-3 175B	175,000	3,100,000

On a single GPU, it would take 335 years to train GPT-3

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

Why Is Extensive Training Is Costly?

- Time-consuming
- Expensive
- Increased environmental impact from extra computations

Extensive Training Is Costly; e.g., Training BERT Cost:

~$80k-$1.6m:

THE COST OF TRAINING NLP MODELS
A CONCISE OVERVIEW

Or Sharir
AI21 Labs
ors@ai21.com

Barak Peleg
AI21 Labs
barakp@ai21.com

Yoav Shoham
AI21 Labs
yoavs@ai21.com

as much energy as a trans-American flight:

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

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

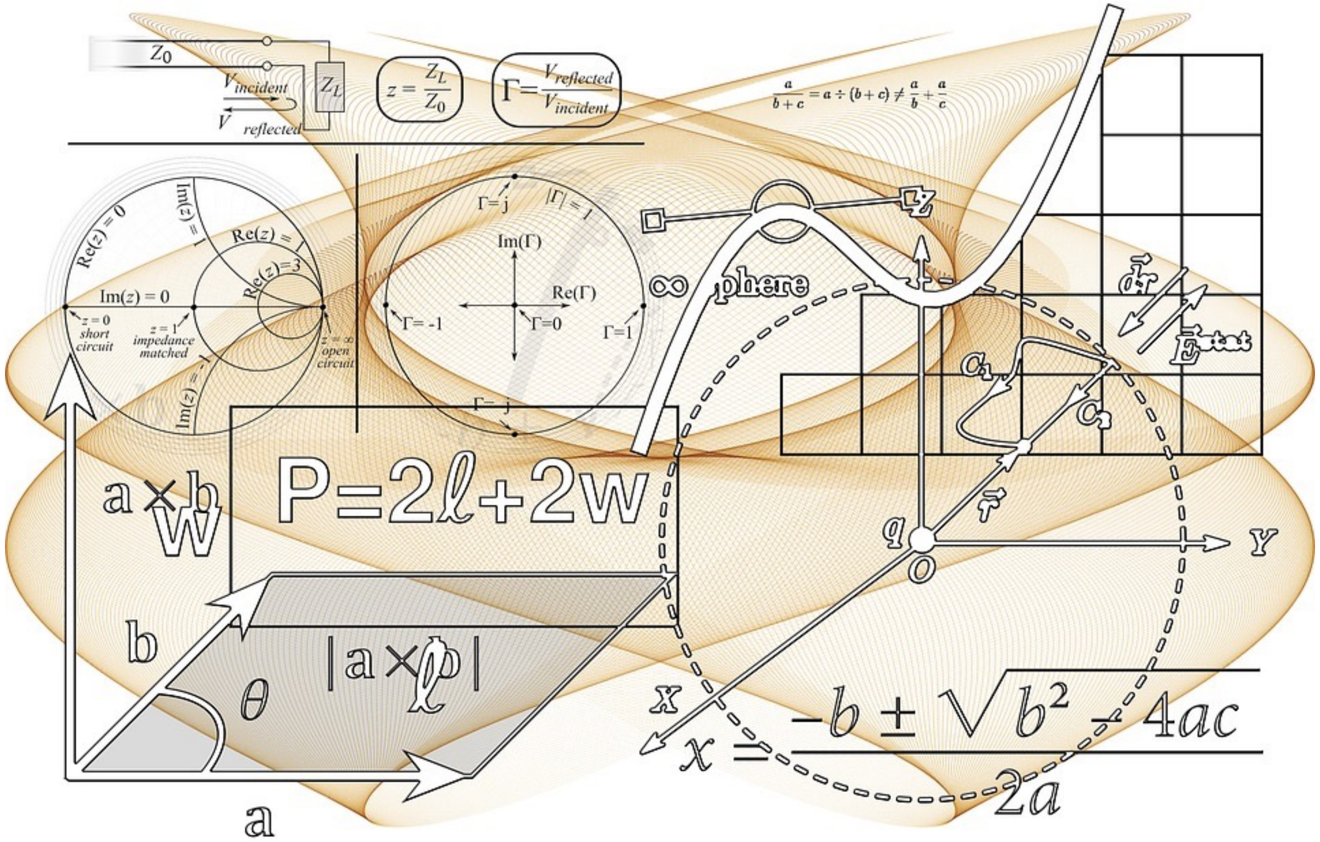
Today's Topics

- Motivation
- Efficient learning: curriculum learning
- Efficient learning: active learning
- Efficient learning: other considerations
- Faculty course questionnaire

How to teach machines to learn more efficiently?

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

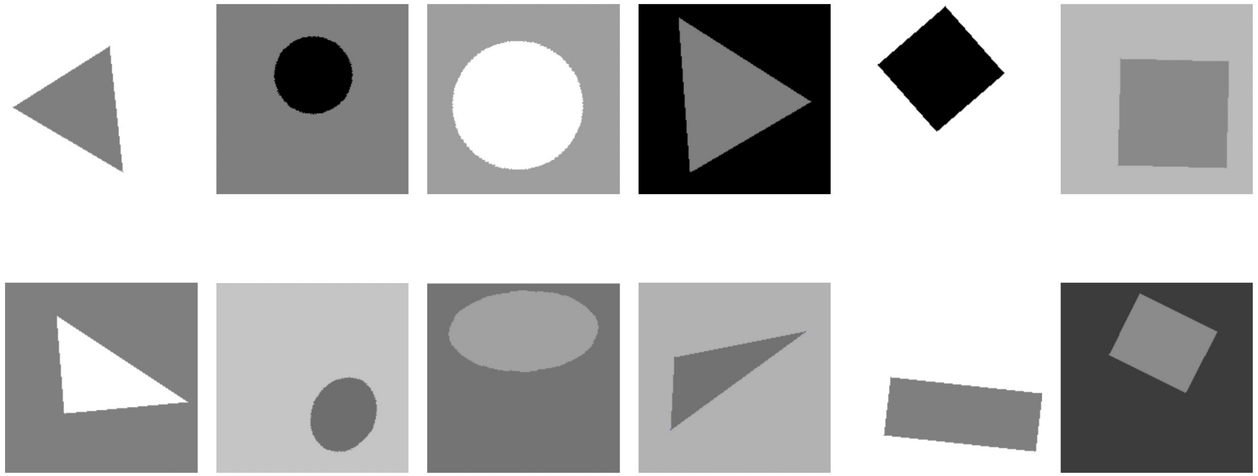
Curriculum

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.

Tasks

1. Classify each shape as rectangle, ellipse, or triangle



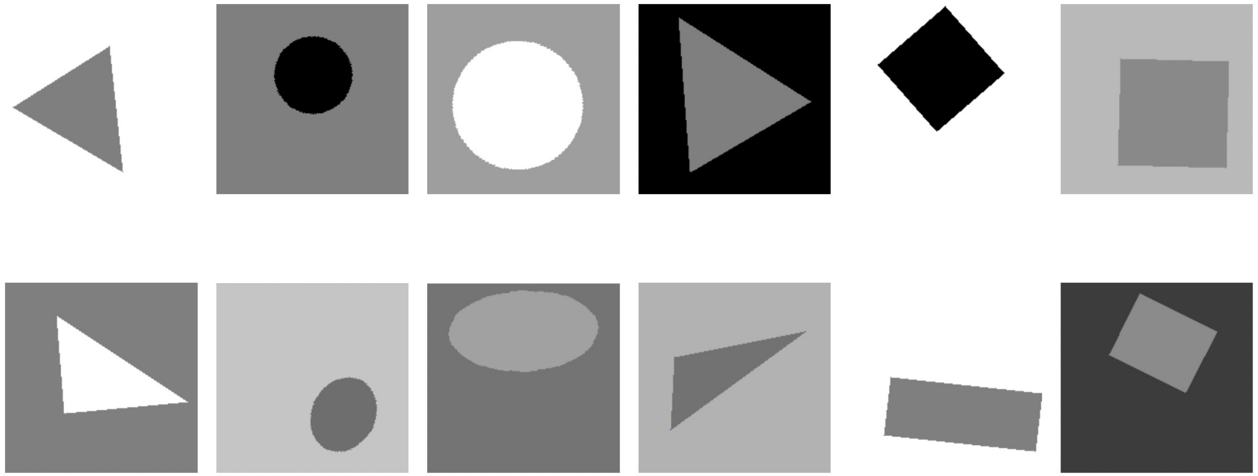
2. Predict the next word

Background music from a ________

Bengio et al., Curriculum Learning, 2009

Shape Prediction: Curriculum Learning

1. Classify each shape as rectangle, ellipse, or triangle



Architecture: 3-layer neural network

Easy (Basic): less shape variability (squares, circles, and equilateral triangles); 10,000 examples

Hard (Geom): more shape variability (rectangles, ellipses, and triangles); 10,000 examples

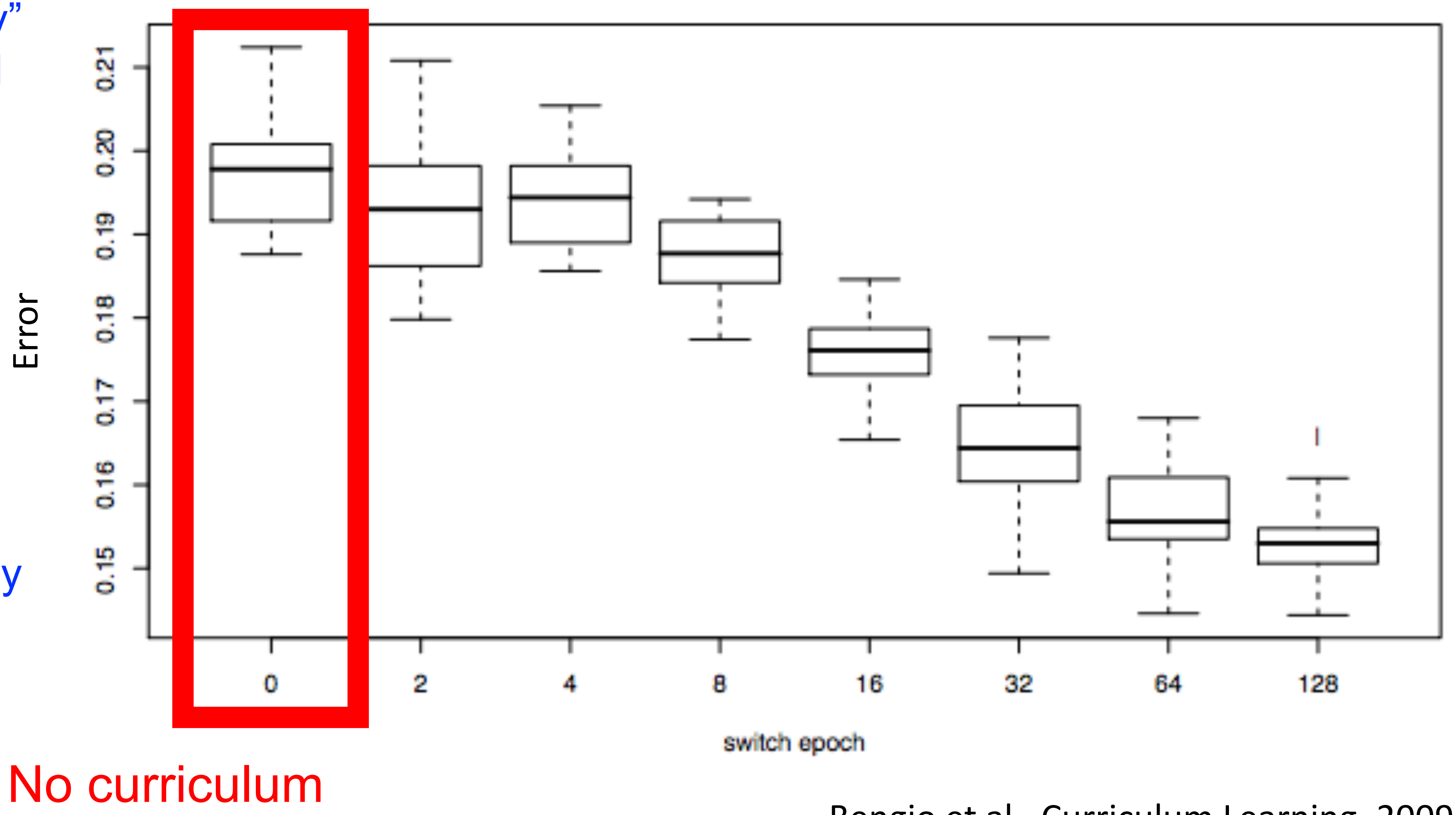
Bengio et al., Curriculum Learning, 2009

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

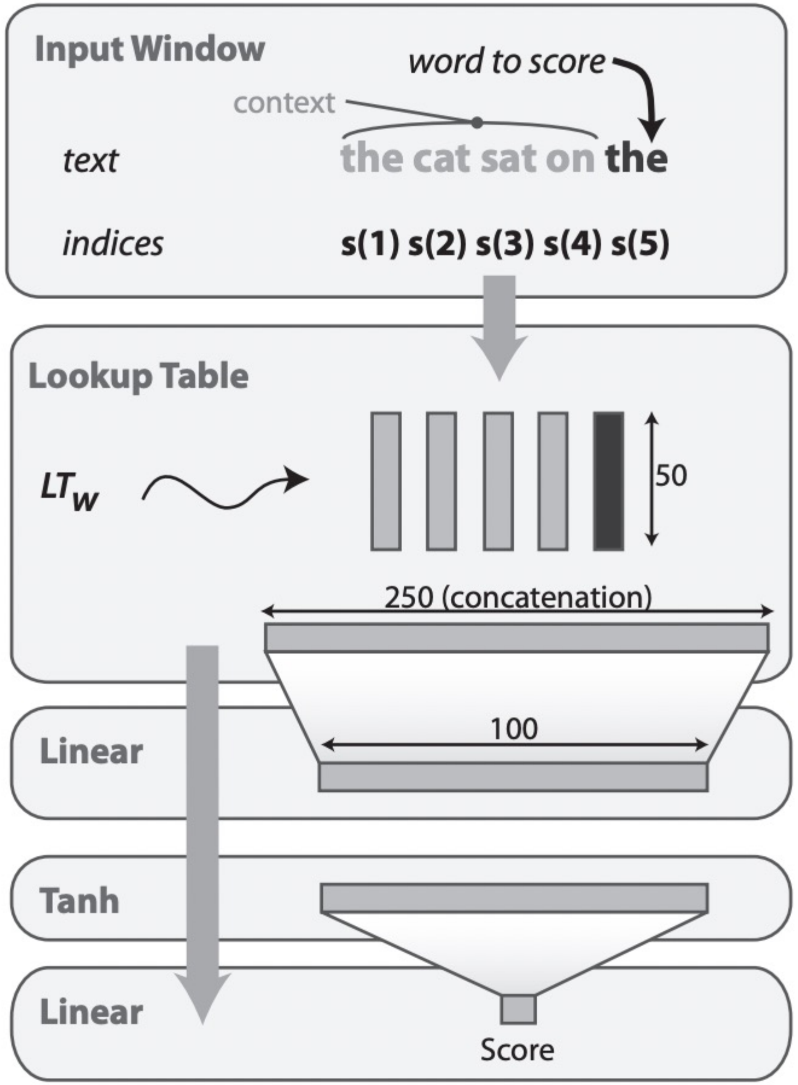
Next Word Prediction: Curriculum Learning

Architecture:
context size
set to 5

Easy: 5,000 most
frequent words

Hard: additional 5,000
words at each epoch
until 20,000 words

Examples with words
not in the vocab were
discarded from training



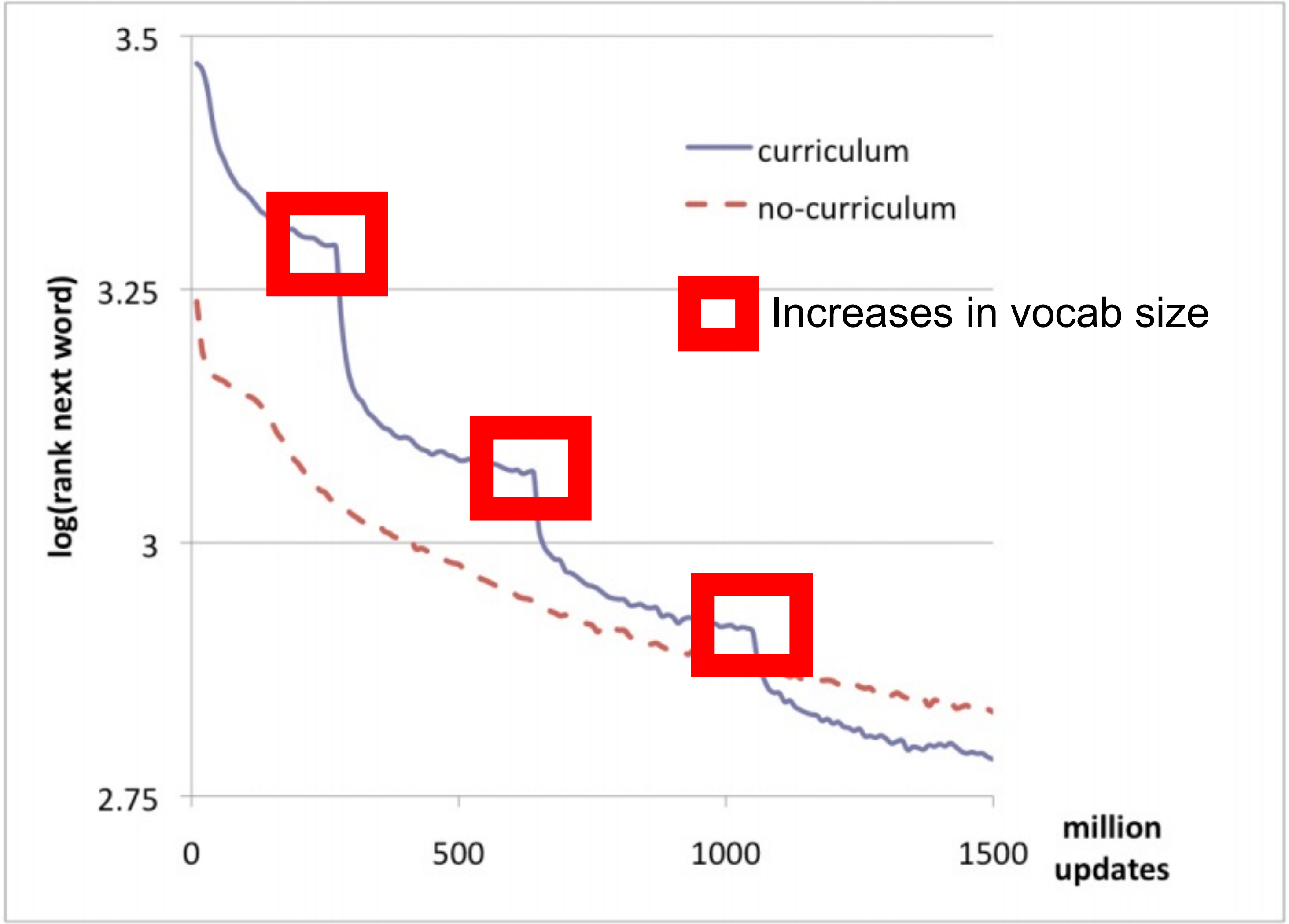
2. Predict the next word

Background music from a ______

Bengio et al., Curriculum Learning, 2009

Next Word Prediction: Curriculum Learning

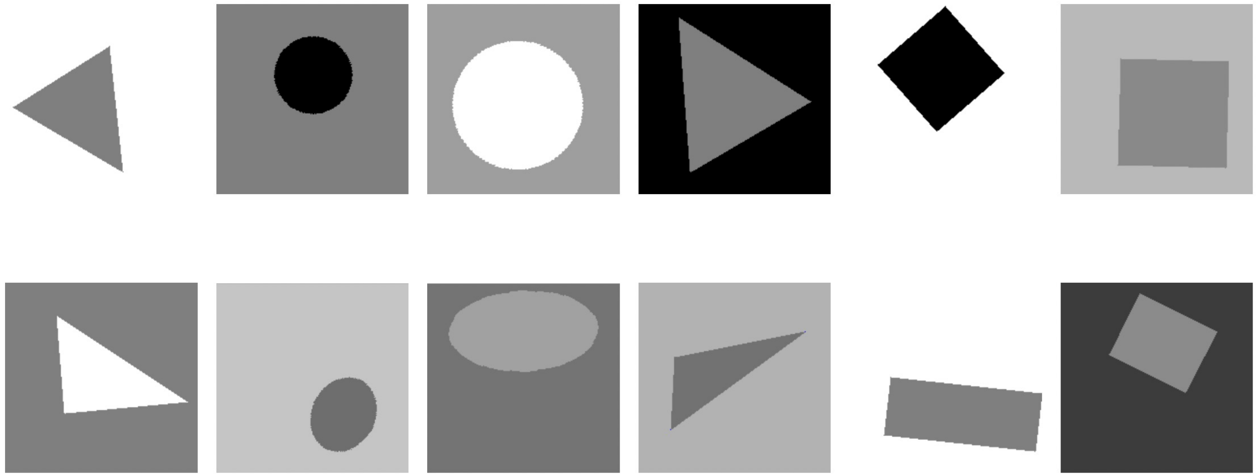
What are benefits of curriculum learning?



Bengio et al., Curriculum Learning, 2009

Summary: Curriculum Learning is a Form of Transfer Learning that Accelerates Optimization

1. Classify each shape as rectangle, ellipse, or triangle

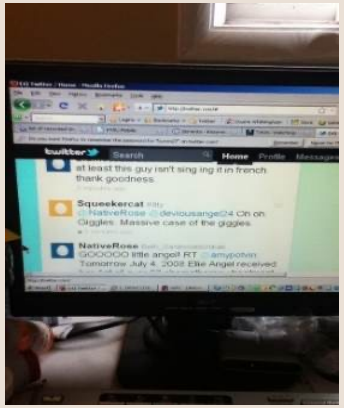


2. Predict the next word

Background music from a ________

Bengio et al., Curriculum Learning, 2009

Key Questions for Curriculum Learning; e.g., for Visual Question Answering



Is my monitor on?

Hi there can you please tell me what flavor this is?

Does this picture look scary?

Which side of the room is the toilet on?

Questions

1. What criteria should be used to order examples?

2. How would you update the training data (and how often)?

Today's Topics

- Motivation
- Efficient learning: curriculum learning
- Efficient learning: active learning
- Efficient learning: other considerations
- Faculty course questionnaire

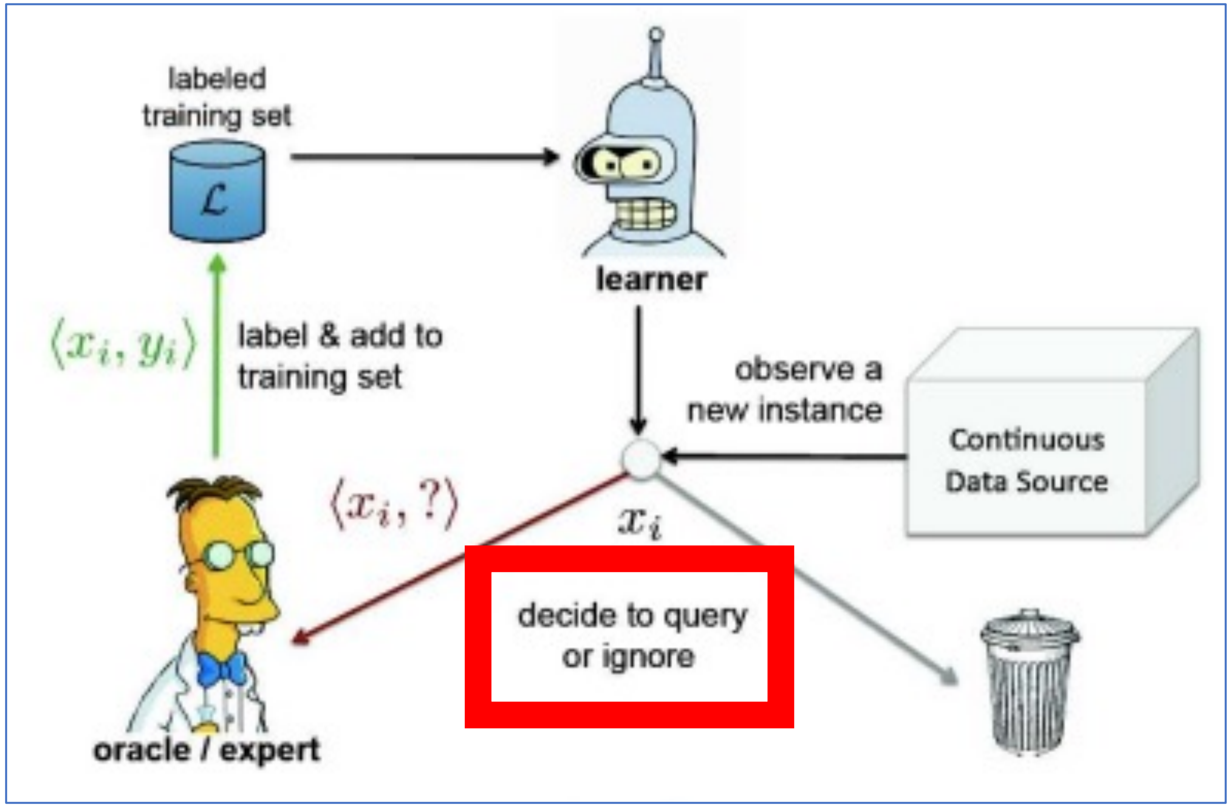
How to teach machines with minimal human supervision?

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

e.g., limited funding

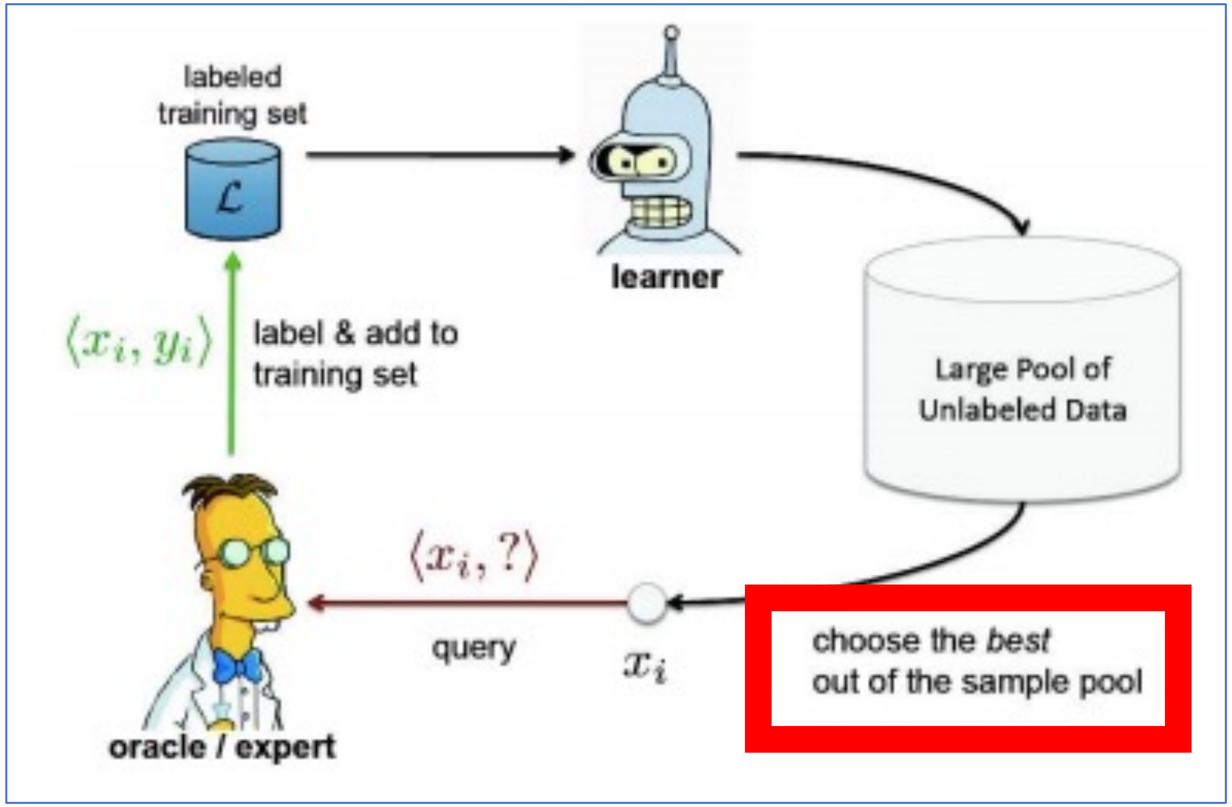
Idea: Choose Most Informative Data to Label

Stream-Based



Consider one example at a time

Pool-Based



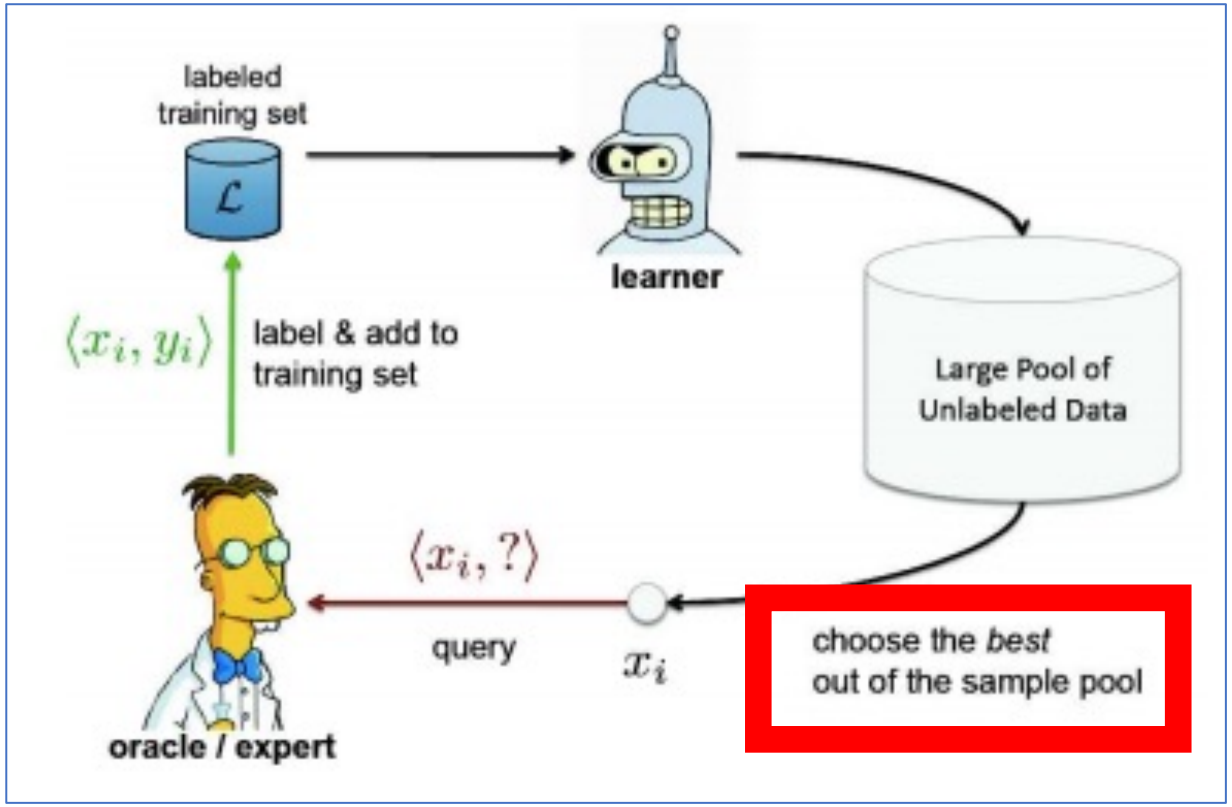
Consider many examples at a time

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

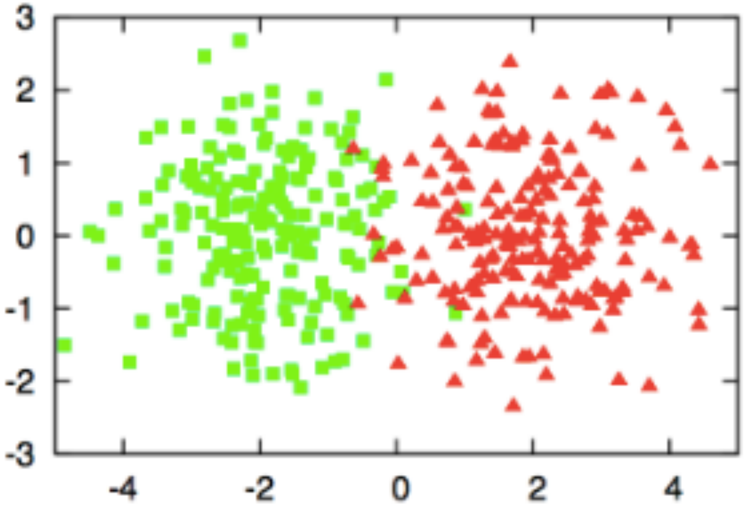
Image Credit: https://www.cs.utah.edu/~piyush/teaching/10-11-slides.pdf

What approach might be effective in identifying the most informative data to label for training neural networks?

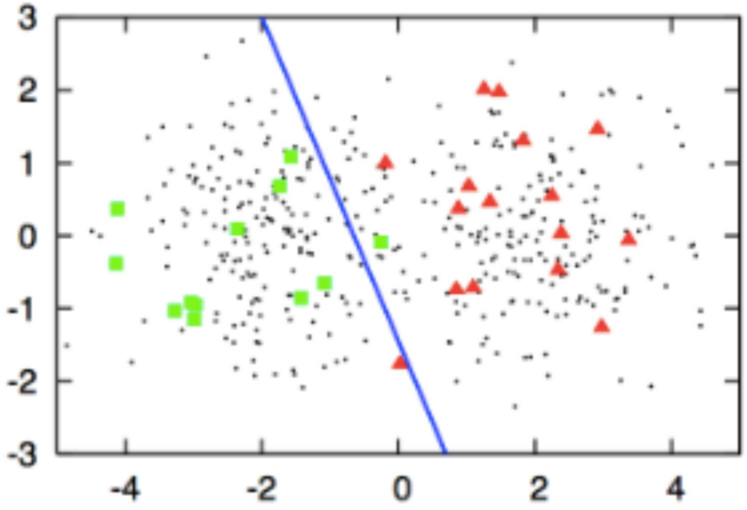
Common Approach: Uncertainty Sampling

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

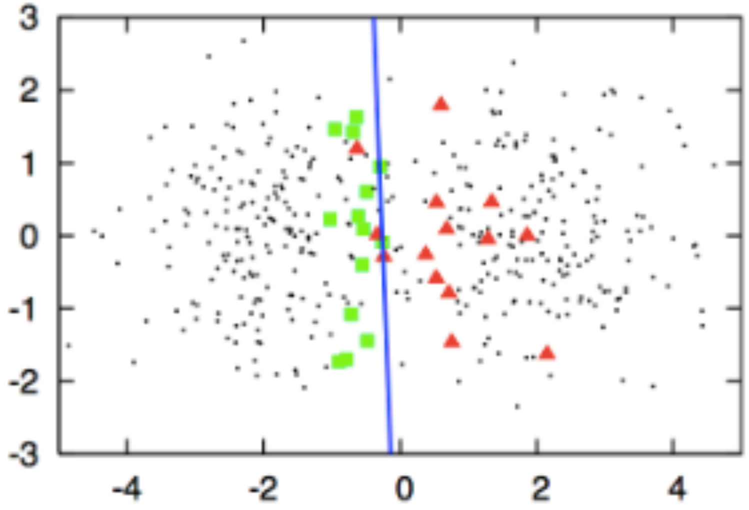
True Representation (Assume Labels Are Not Known)



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.,

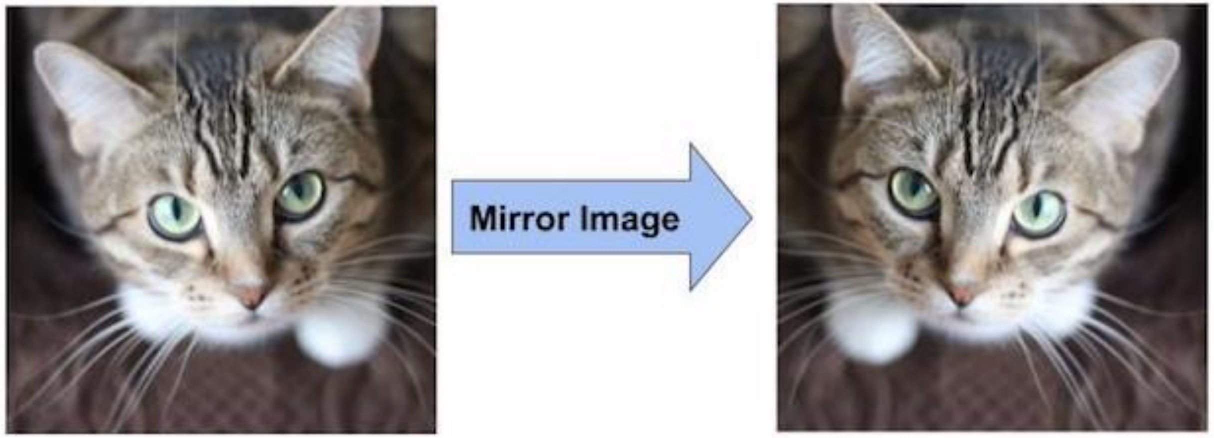


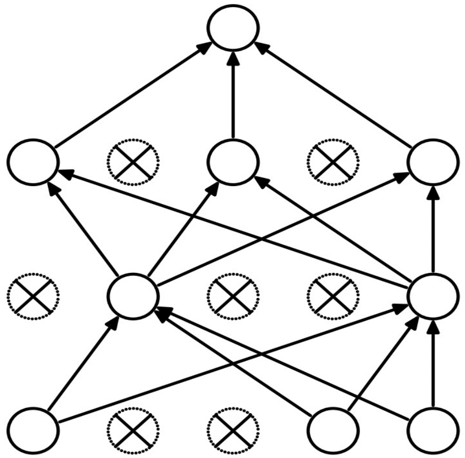
Figure Source: https://learnopencv.com/understanding-alexnet/

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

2. Multiple neural networks



(b) After applying dropout.

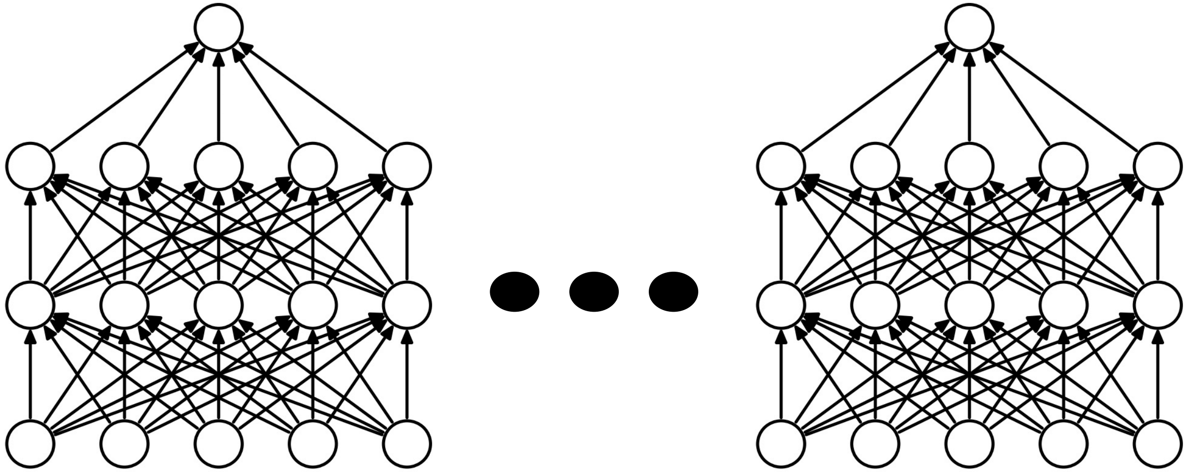


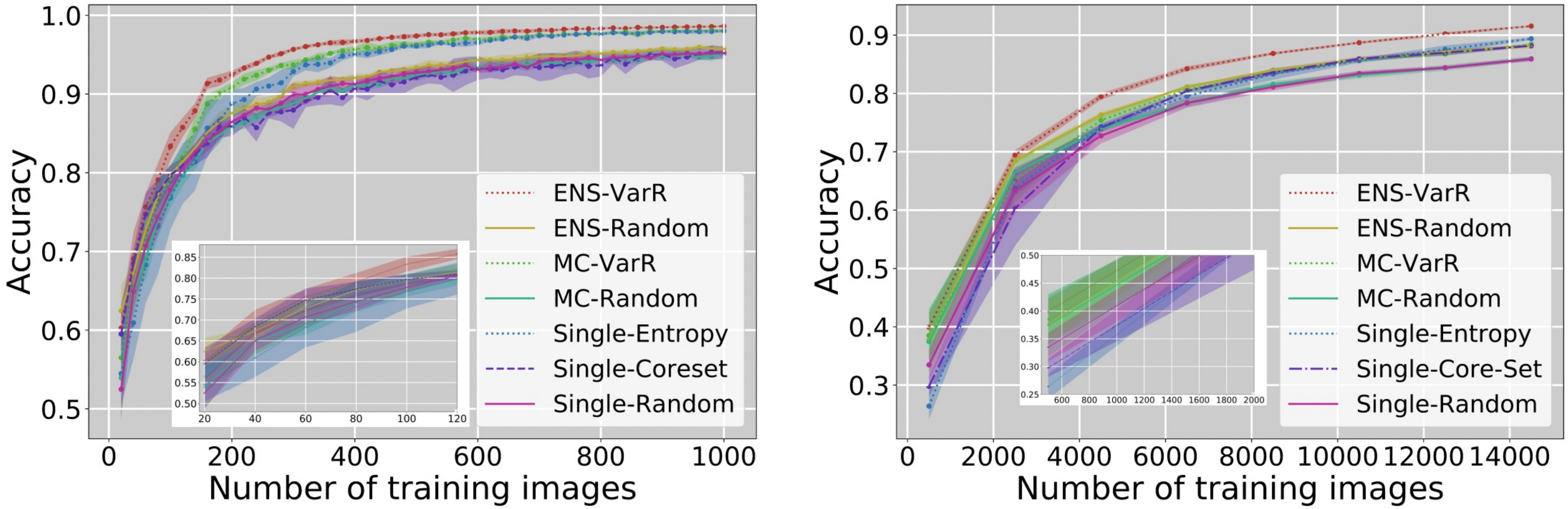
Figure Source: Srivastava et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research. 2014

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

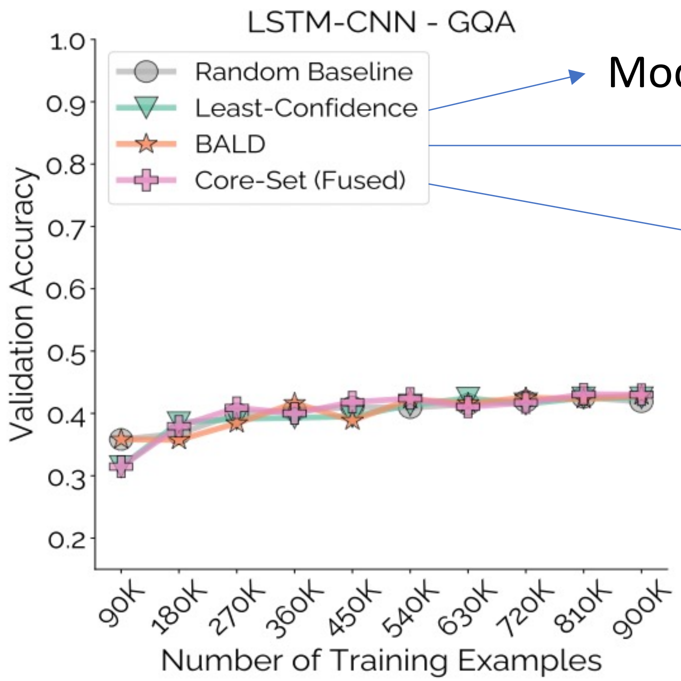


(a) MNIST on S-CNN (b) CIFAR-10 on DenseNet

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

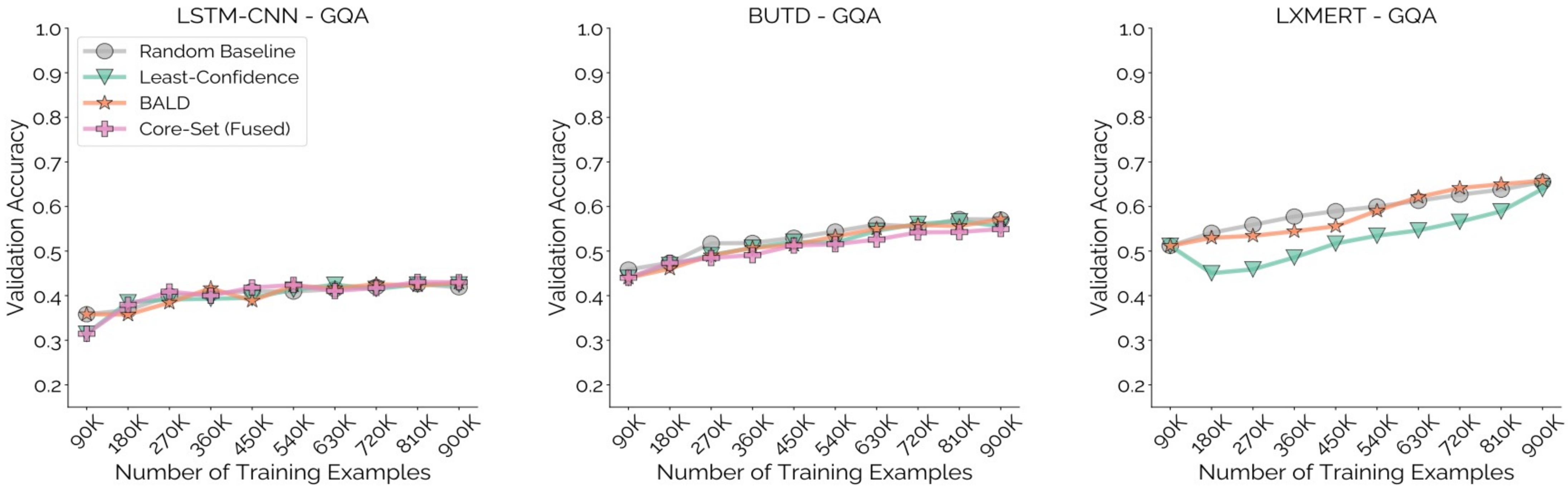
Based on ensemble from dropout

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

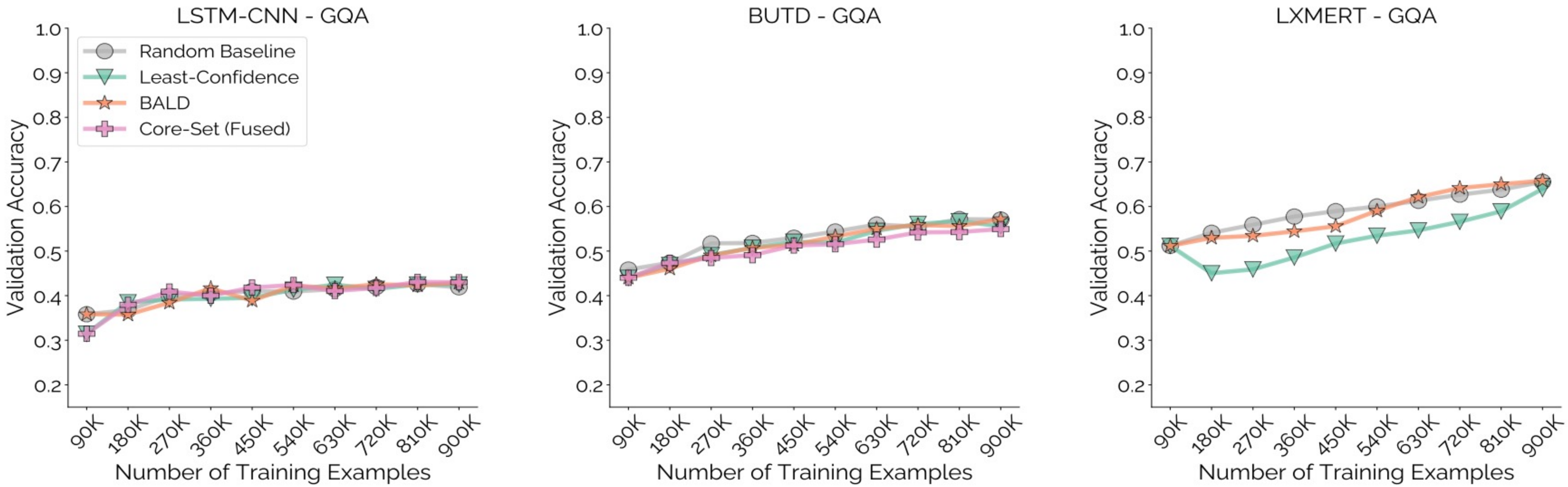
How do the 3 AL methods compare to random selection for the 3 VQA models?



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?



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?

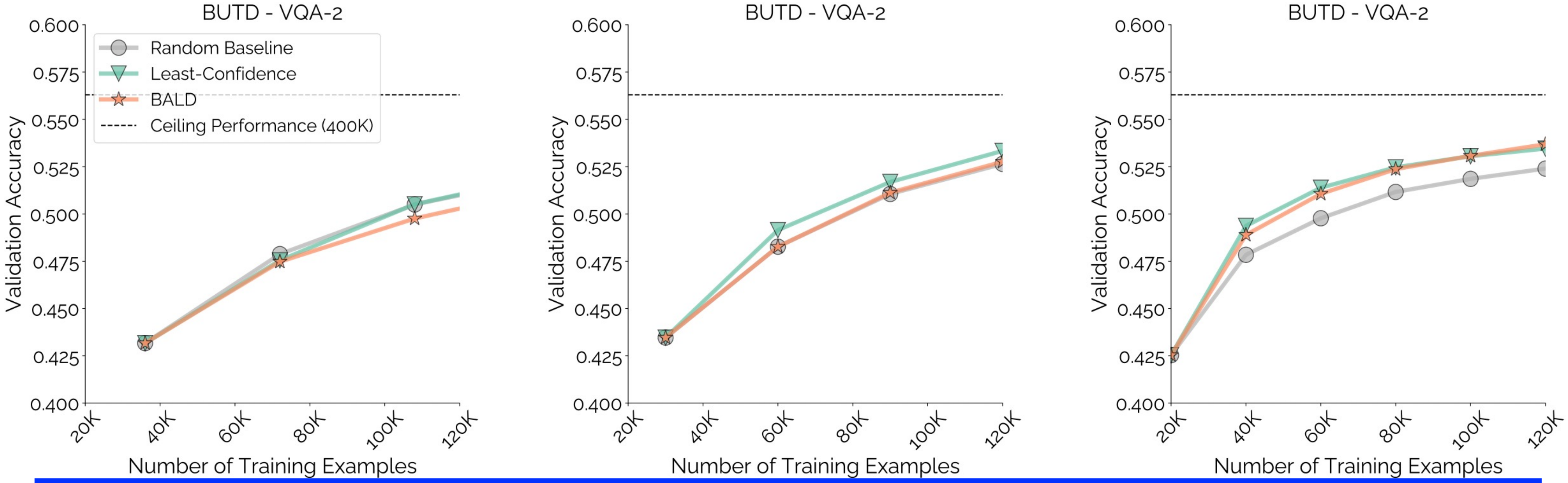
- Challenging examples to learn are sampled; e.g.,

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

What is the performance trend for AL approaches compared to random selection when removing "challenging" examples from data pool?



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

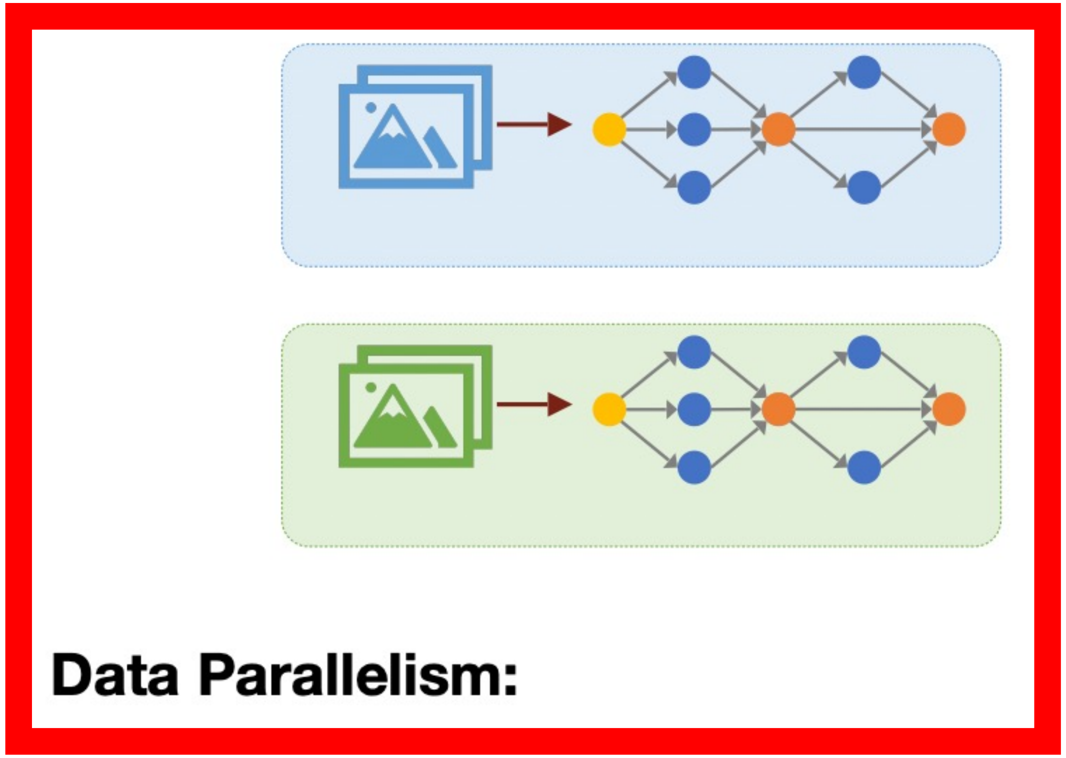
Karamcheti et al. Mind your outliers! Investigating the negative impact of outliers on active learning for visual question answering. Association for Computational Linguistics (ACL) 2021

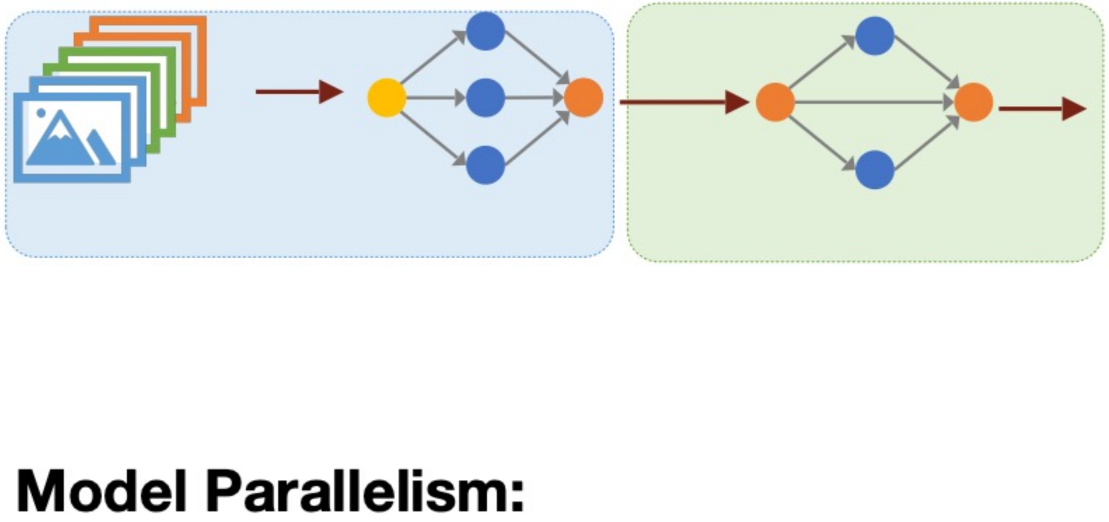
Today’s Topics

- Motivation
- Efficient learning: curriculum learning
- Efficient learning: active learning
- Efficient learning: other considerations
- Faculty course questionnaire

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

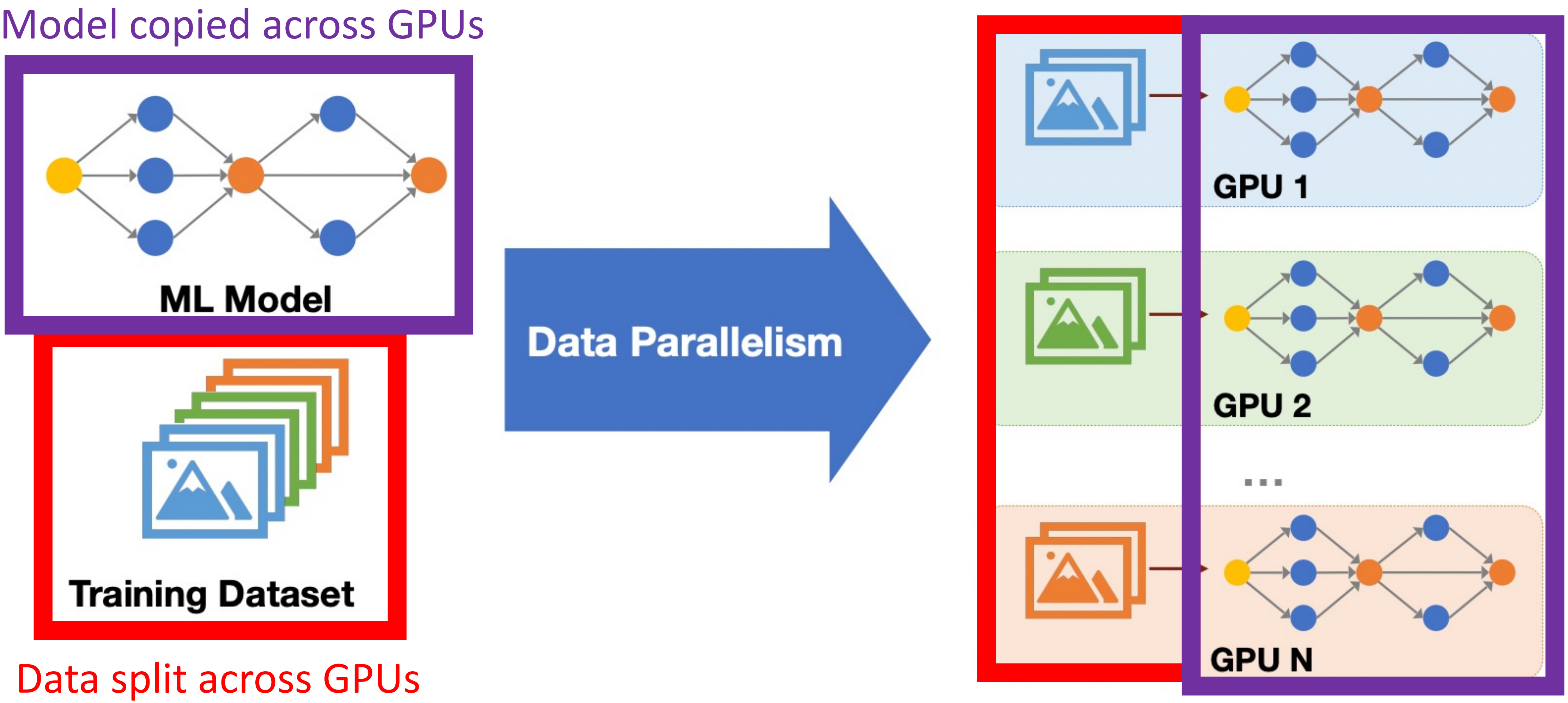
Distributed Training





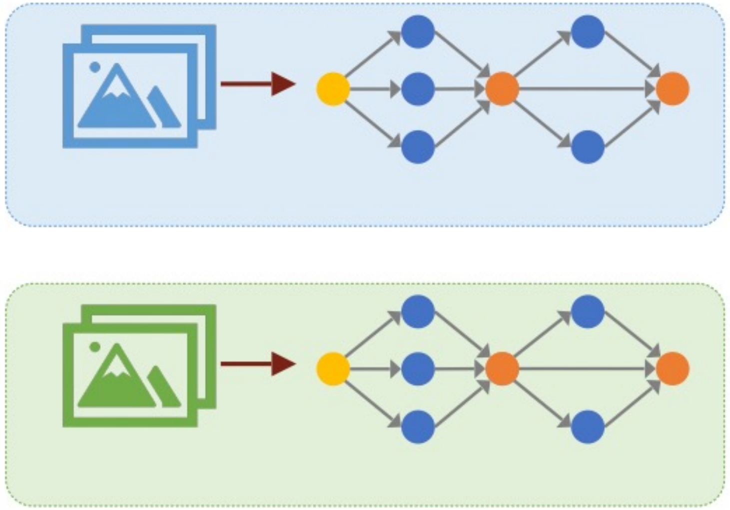
https://hanlab.mit.edu/files/course/slides/MIT-TinyML-Lec13-Distributed-Training-I.pdf

Distributed Training: Data Parallelism

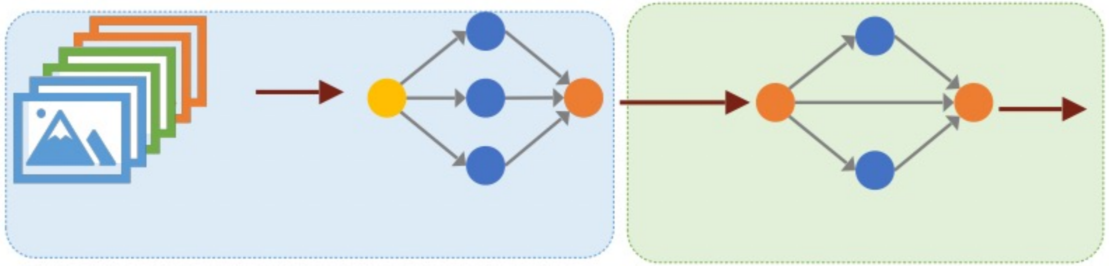


https://hanlab.mit.edu/files/course/slides/MIT-TinyML-Lec13-Distributed-Training-I.pdf

Distributed Training



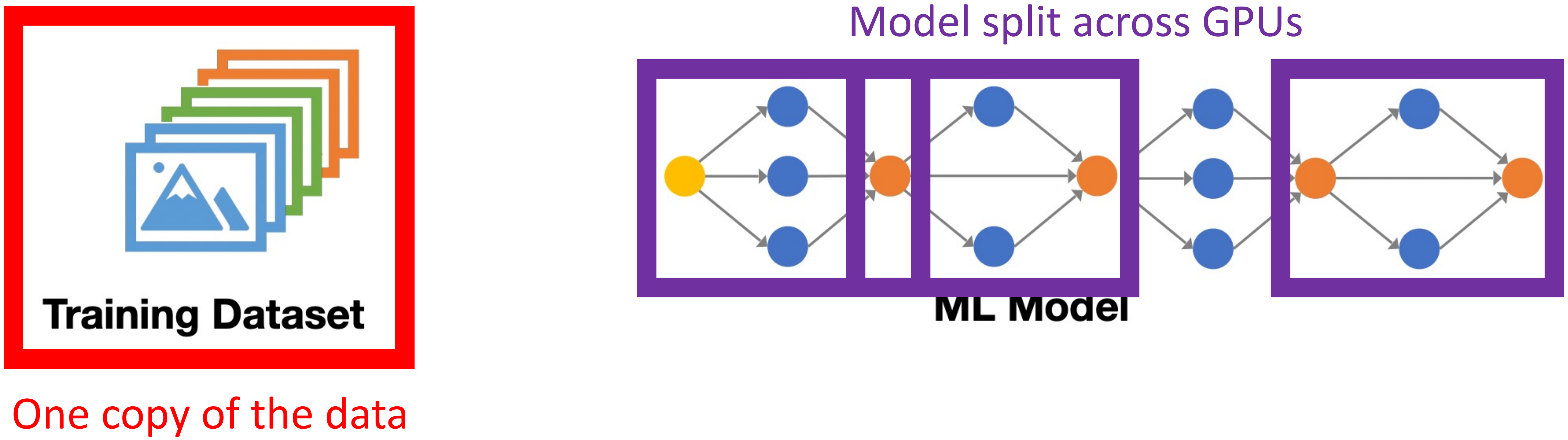
Data Parallelism:

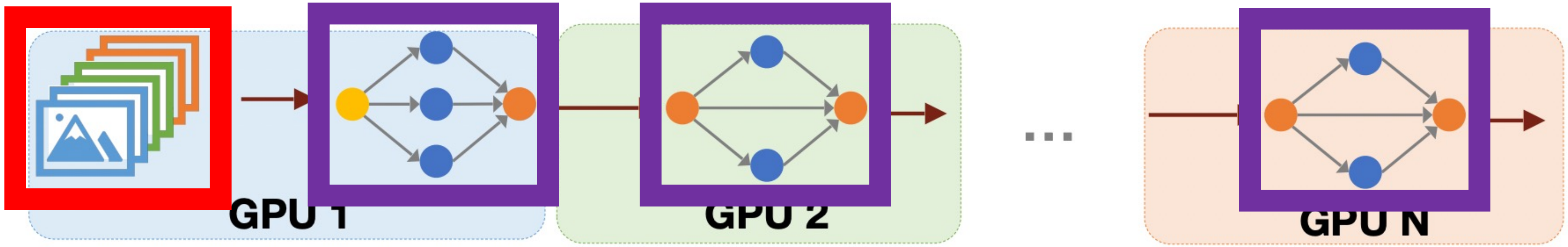


Model Parallelism:

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

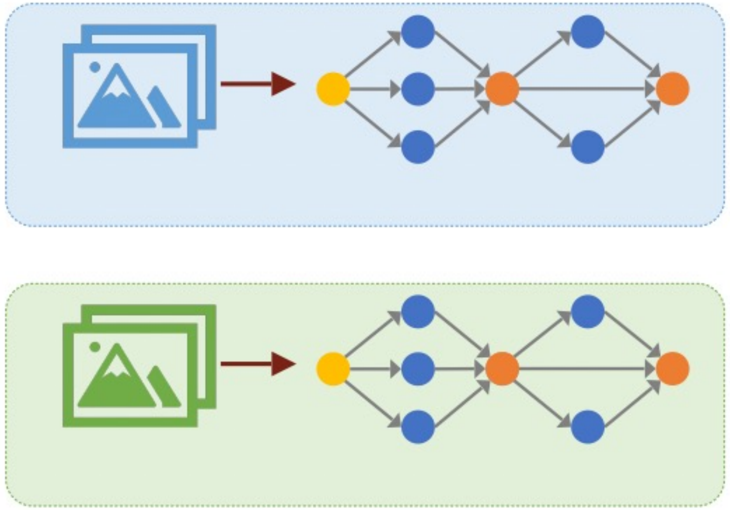
Distributed Training: Model Parallelism

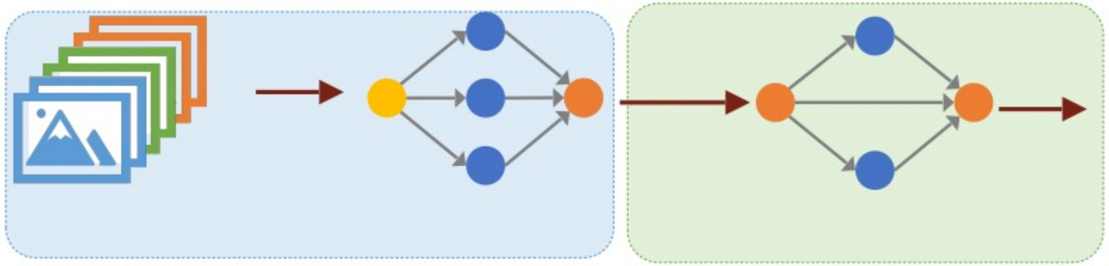




https://hanlab.mit.edu/files/course/slides/MIT-TinyML-Lec13-Distributed-Training-I.pdf

Distributed Training





Data Parallelism:

- Split the data
- Same model across devices
- Easy to parallelize, high utilization
- N copies of model

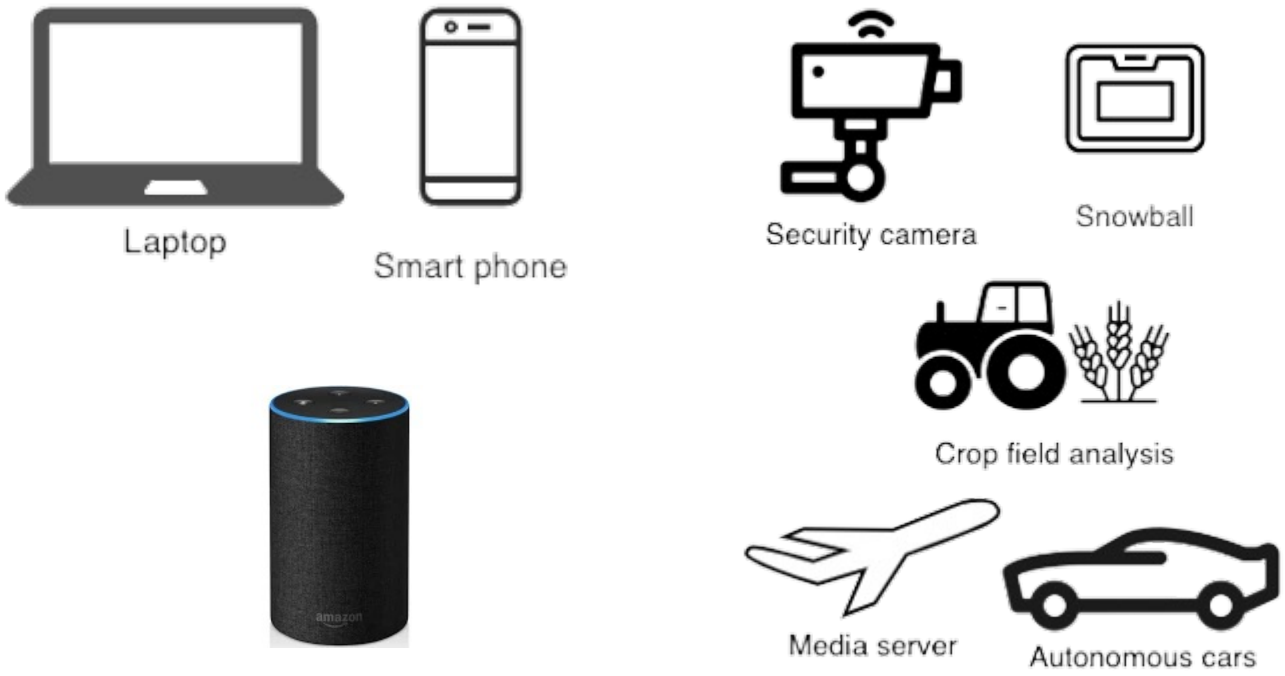
Model Parallelism:

- Split the model
- Move activations through devices
- Hard to parallelize, load balancing issue
- Single copy of model

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

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

On-Device Training: Learn with Limited Memory and Compute



DL systems may need to adapt to users' data on-device for reasons such as poor/no internet connection and privacy

Tutorial: https://hanlab.mit.edu/files/course/slides/MIT-TinyML-Lec15-On-Device-Training-And-Transfer-Learning-I.pdf
Figure: https://aws.amazon.com/blogs/machine-learning/demystifying-machine-learning-at-the-edge-through-real-use-cases/

Today's Topics

- Motivation
- Efficient learning: curriculum learning
- Efficient learning: active learning
- Efficient learning: other considerations
- Faculty course questionnaire: https://colorado.campuslabs.com/courseeval

The End