Efficient Computer Vision

Danna Gurari
University of Colorado Boulder
Fall 2024

Review

- Previous lectures:
  - Student-led lectures
- Assignments:
  - Project presentation poster due in 1 week
  - Project presentation due in 1.5 weeks
  - Peer evaluation due in 1.5 weeks (in-class activity)
  - Project report due in 2.5 weeks
- Questions?

Efficient Computer Vision

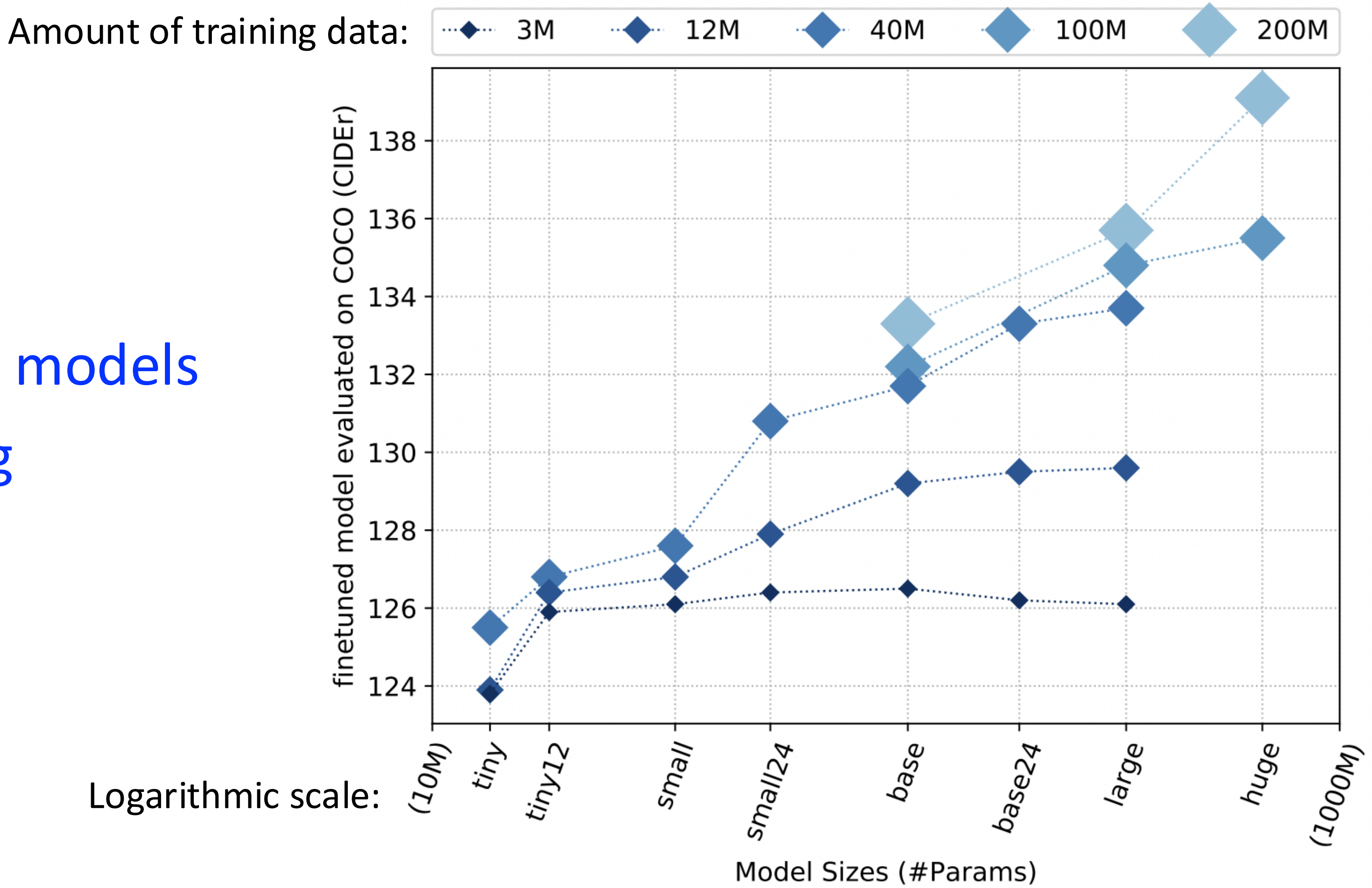
- Motivation
- Model Compression
- Curriculum Learning
- Active Learning
- Faculty Course Questionnaire (FCQ)

Efficient Computer Vision

- Motivation
- Model Compression
- Curriculum Learning
- Active Learning
- Faculty Course Questionnaire (FCQ)

What Are Common Trends for CV Models?

1. Parameter-heavy models
2. Extensive training



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

Trend: Parameter-Heavy Models

How many parameters are estimated to be in GPT-4 (was used for ChatGPT)?

(a) 176 million

(b) 1.76 billion

(c) 17.6 billion

(d) 170.6 billion

(e) 1.76 trillion

https://en.wikipedia.org/wiki/GPT-4

Trend: Extensive Training

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

(a) 3 million

(b) 30 million

(c) 300 million

(d) 3 billion

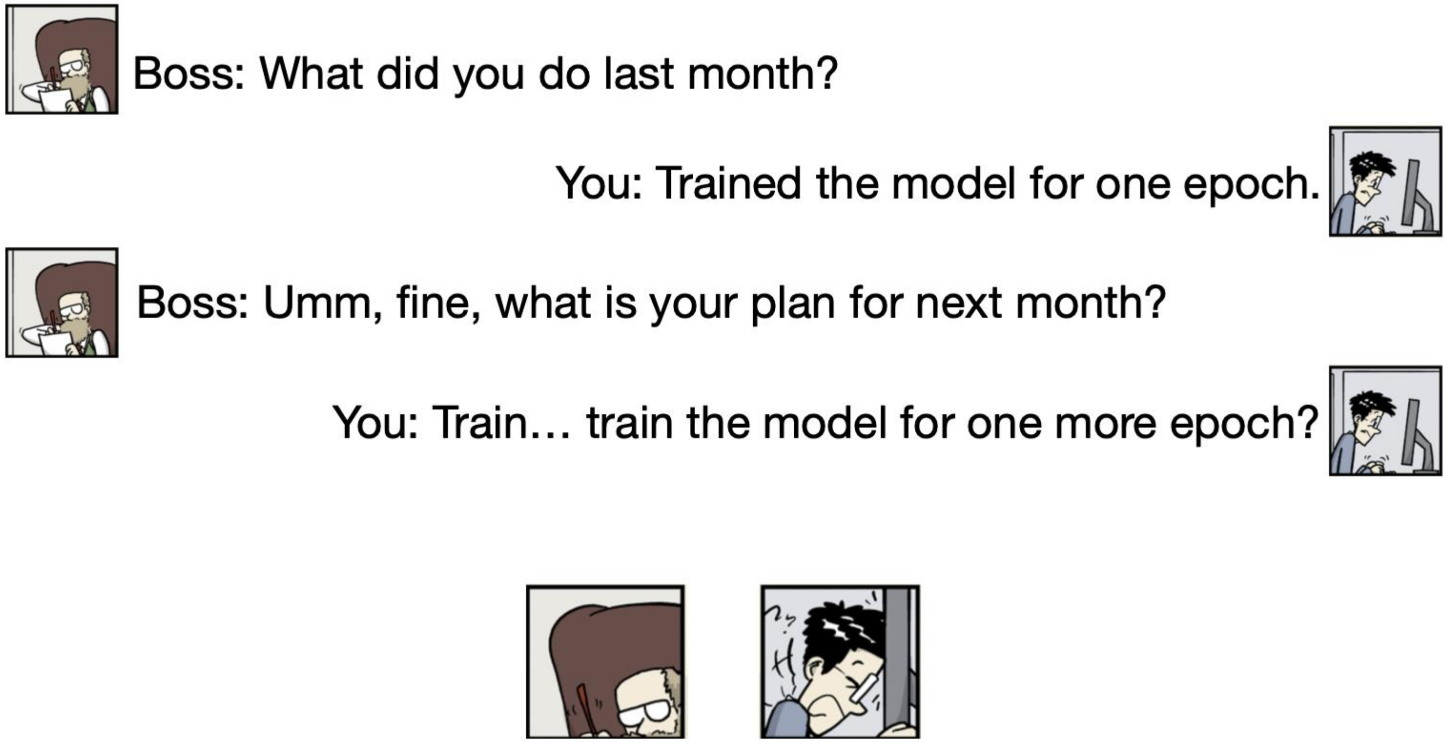
(e) 30 billion

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

Zhai et al. Scaling Vision Transformers. CVPR 2022

Modern Neural Networks Are a Mismatch for Many Real-World Applications

- Time-consuming (e.g., incompatible for real-time applications)



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

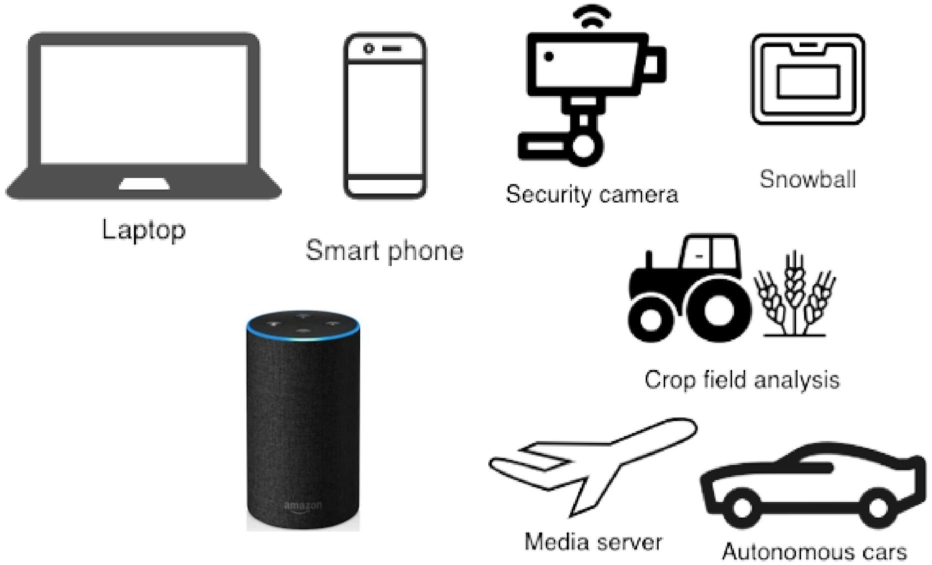
Modern Neural Networks Are a Mismatch for Many Real-World Applications

- Time-consuming (e.g., incompatible for real-time applications)
- Large memory footprint (e.g., incompatible with edge devices)

https://www.ephotozine.com/article/19-things-to-look-out-for-in-a-smartphone-camera--31055

https://en.wikipedia.org/wiki/Wearable_technology

https://www.buzzfeednews.com/article/katienotopoulos/facebook-is-making-camera-glasses-ha-ha-oh-no



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

Modern Neural Networks Are a Mismatch for Many Real-World Applications

- Time-consuming (e.g., incompatible for real-time applications)

- Large memory footprint (e.g., incompatible with edge devices)

- Large computational cost (e.g., large environmental costs)

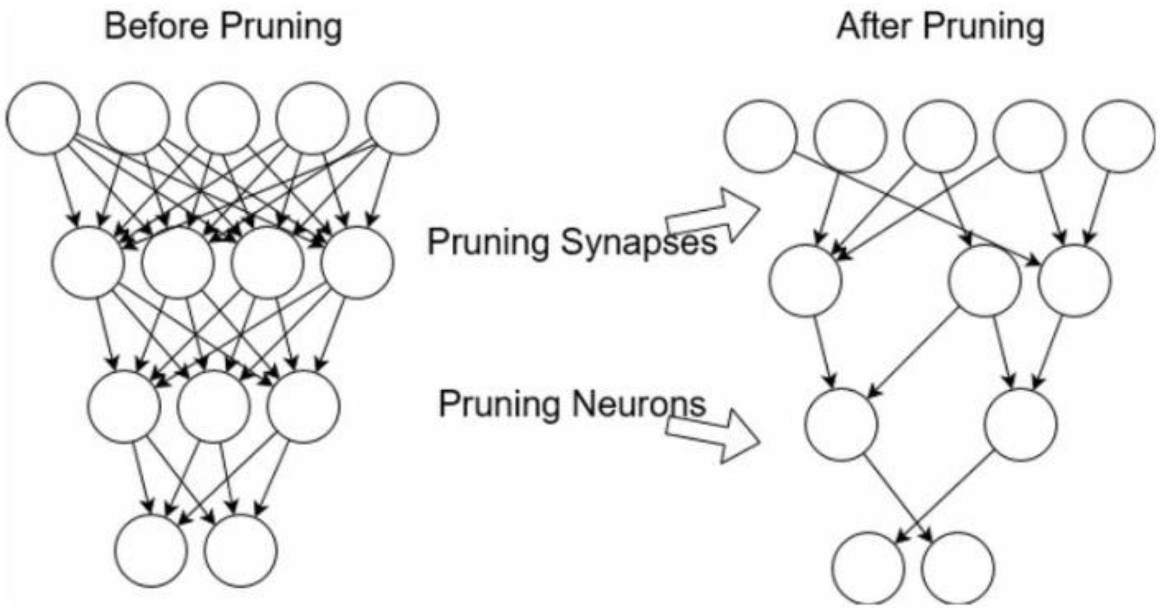
Idea: develop models that are more compact and learn more efficiently (i.e., faster and with less data)

Efficient Computer Vision

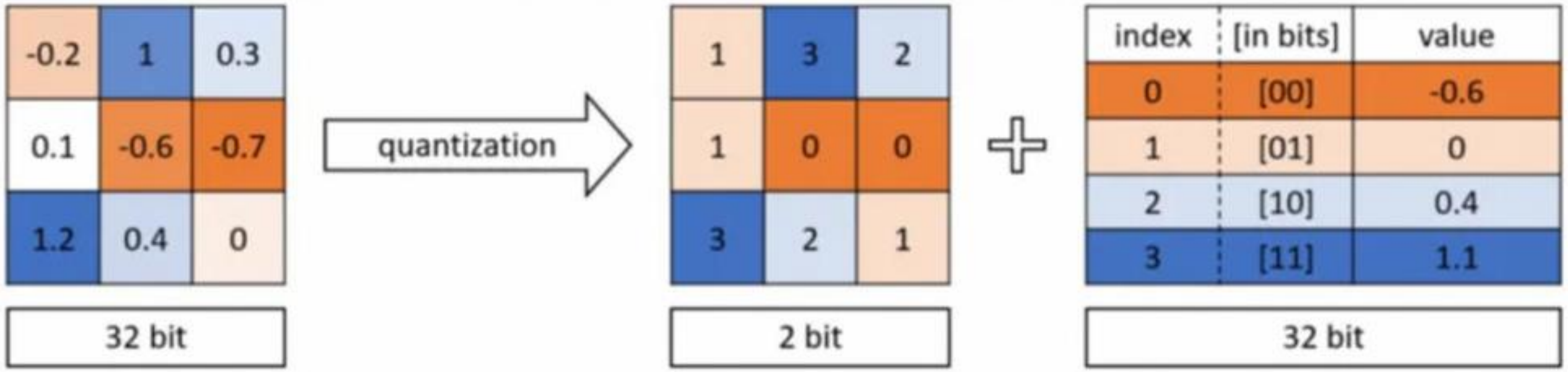
- Motivation
- Model Compression
- Curriculum Learning
- Active Learning
- Faculty Course Questionnaire (FCQ)

Dated Compression Approaches

1. Prune networks



2. Represent weights with fewer bits



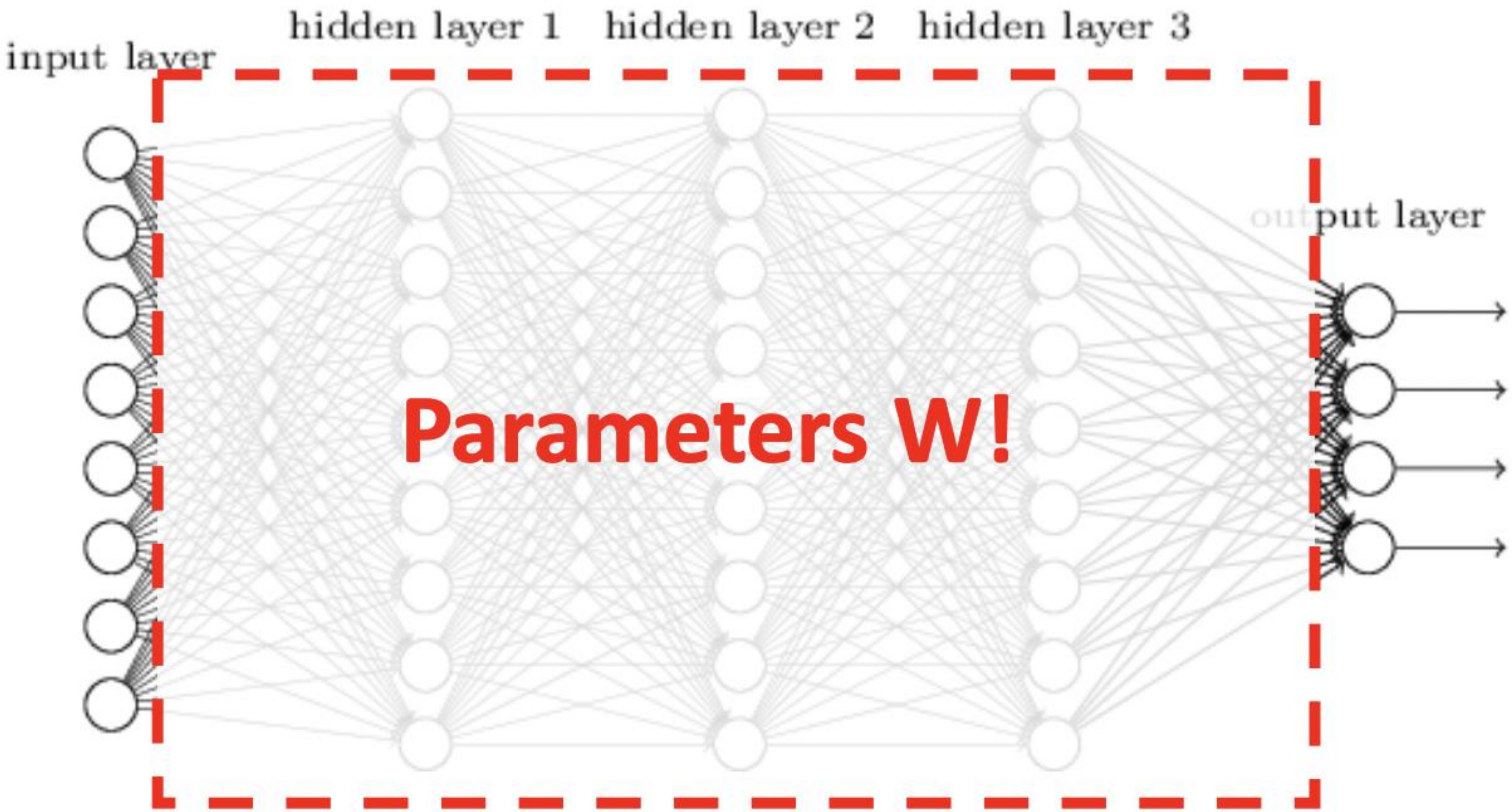
https://xailient.com/blog/4-popular-model-compression-techniques-explained/

Today’s Popular Approach: Knowledge Distillation

A student learns from a knowledgeable teacher

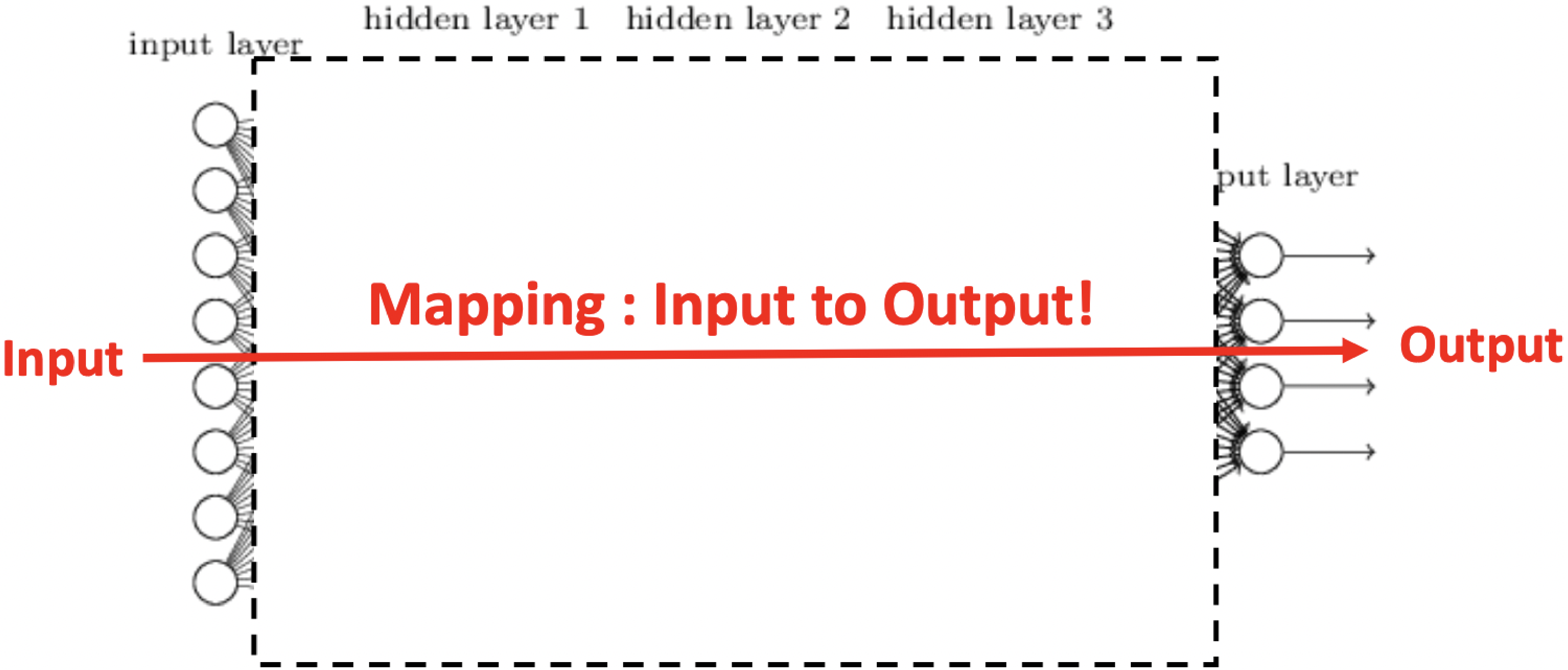
Image source: https://www.waterford.org/education/teacher-student-relationships/

Key Question: What is Knowledge?



http://ir.hit.edu.cn/~xiachongfeng/slides/Knowledge%20Distillation.pdf

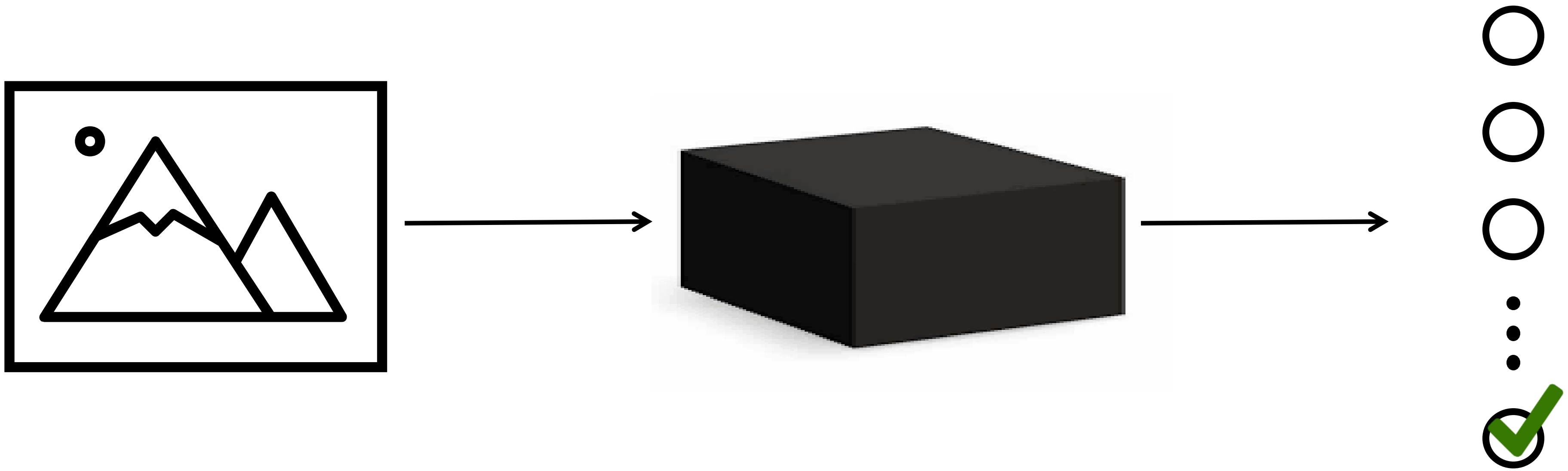
Knowledge Is: Input to Output Mapping



http://ir.hit.edu.cn/~xiachongfeng/slides/Knowledge%20Distillation.pdf

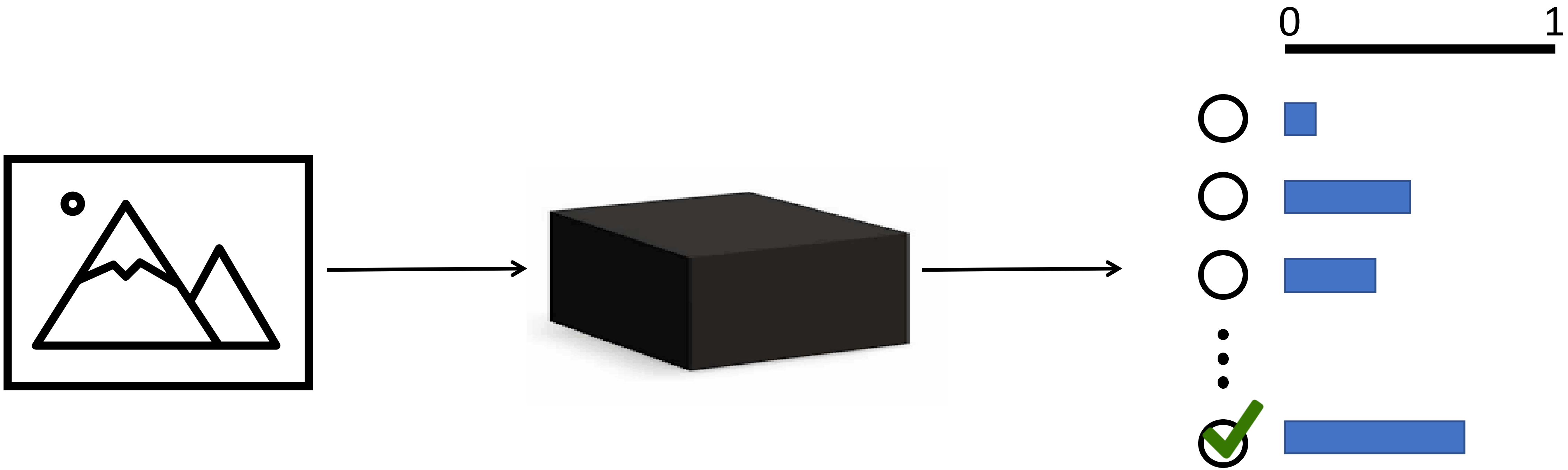
Knowledge Is: Input to Output Mapping

Target mapping: ground truth (1-hot vector)



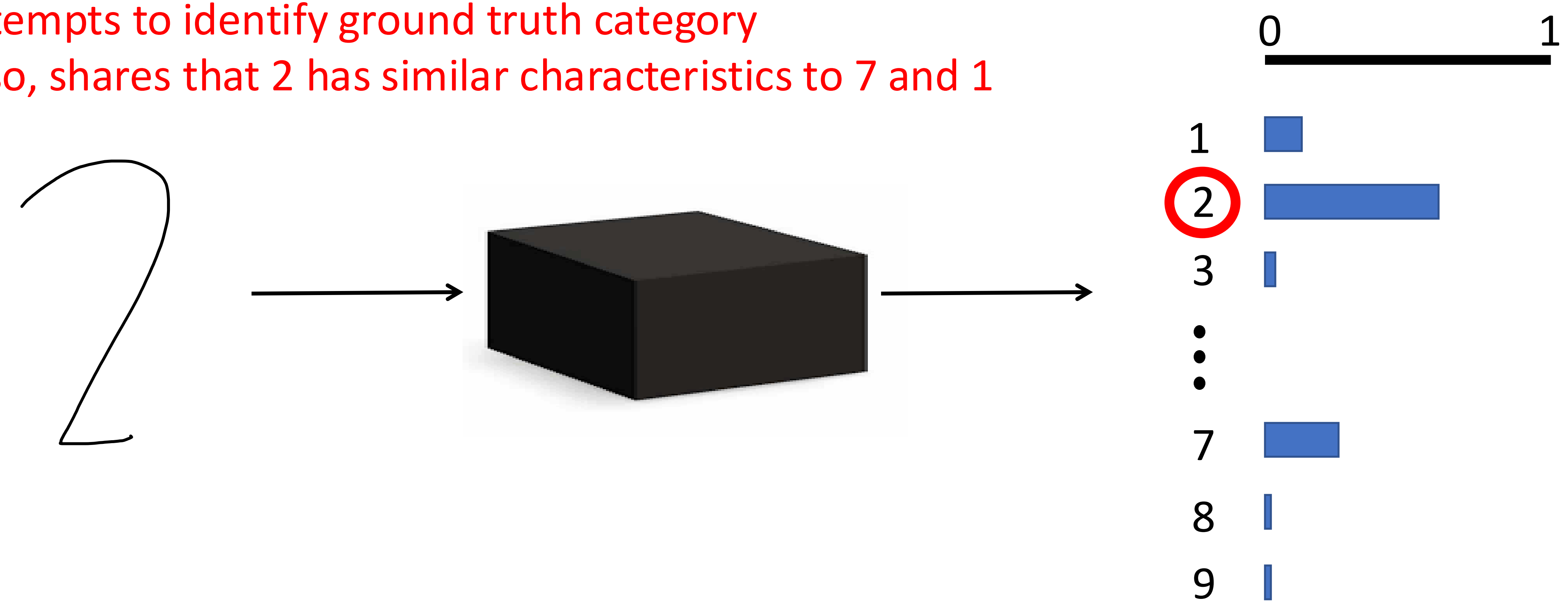
Knowledge Is: Input to Output Mapping

Target mapping: probability distribution from a model offers
further insights into similarities and differences of categories



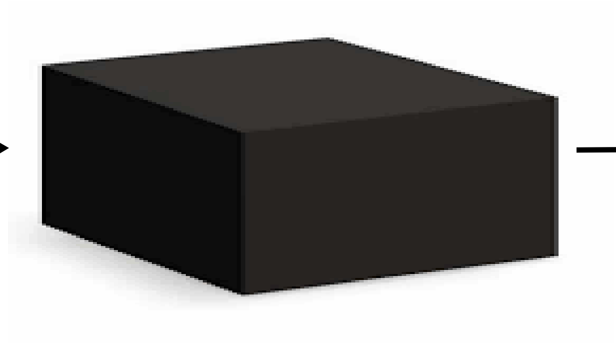
Knowledge Is: Input to Output Mapping

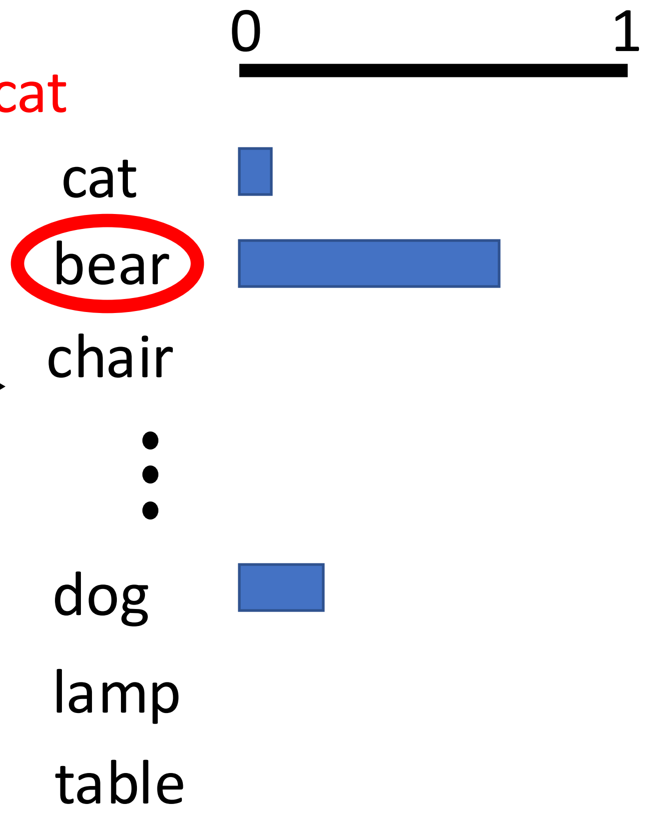
Target mapping: probability distribution from a model offers
further insights into similarities and differences of categories
- Attempts to identify ground truth category
- Also, shares that 2 has similar characteristics to 7 and 1



Knowledge Is: Input to Output Mapping

Target mapping: probability distribution from a model offers
further insights into similarities and differences of categories
- Attempts to identify ground truth category
- Also, shares that bear has similar characteristics to dog and cat





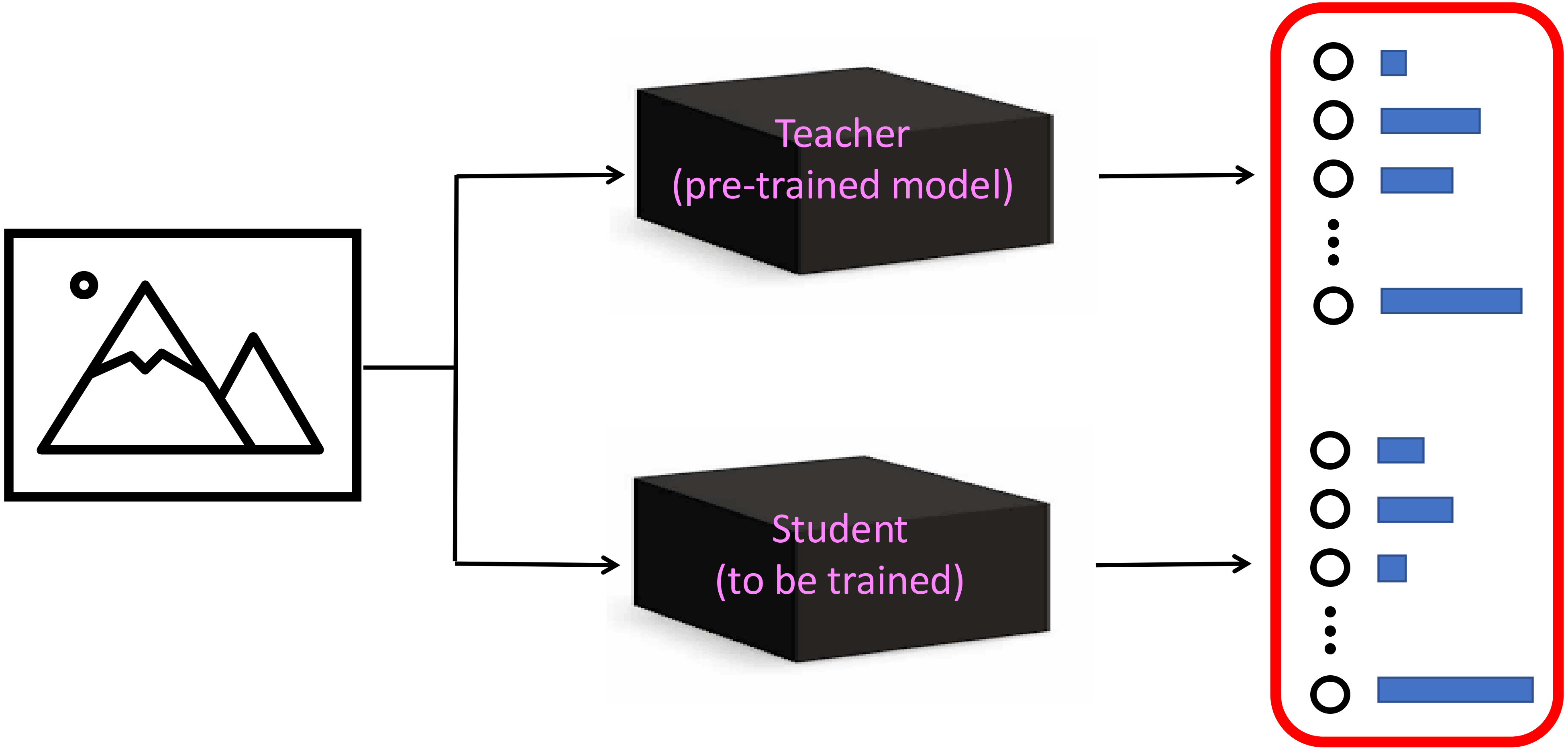
Knowledge Is: Input to Output Mapping

Target mapping: probability distribution from a model offers further insights into similarities and differences of categories
- Attempts to identify ground truth category
- Also, shares that bear has similar characteristics to dog and cat

Idea: teach about ground truth and its relationships to other categories

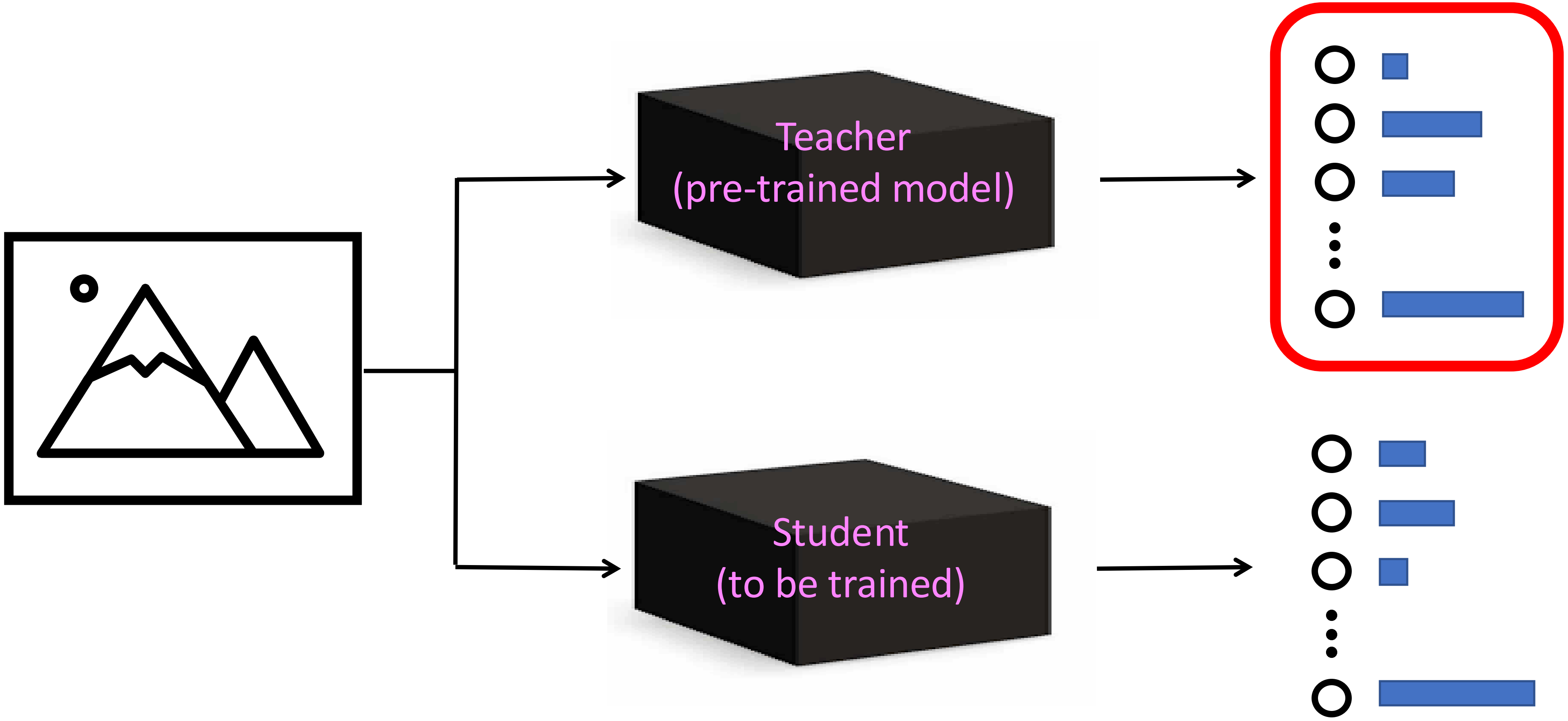
Hinton, Vinyals, and Dean. Distilling the knowledge in a neural network. *arXiv* 2015.

Knowledge Distillation: Teach Student the "Dark Knowledge" of Teacher



Loss computed to bring student's distribution closer to that of the teacher

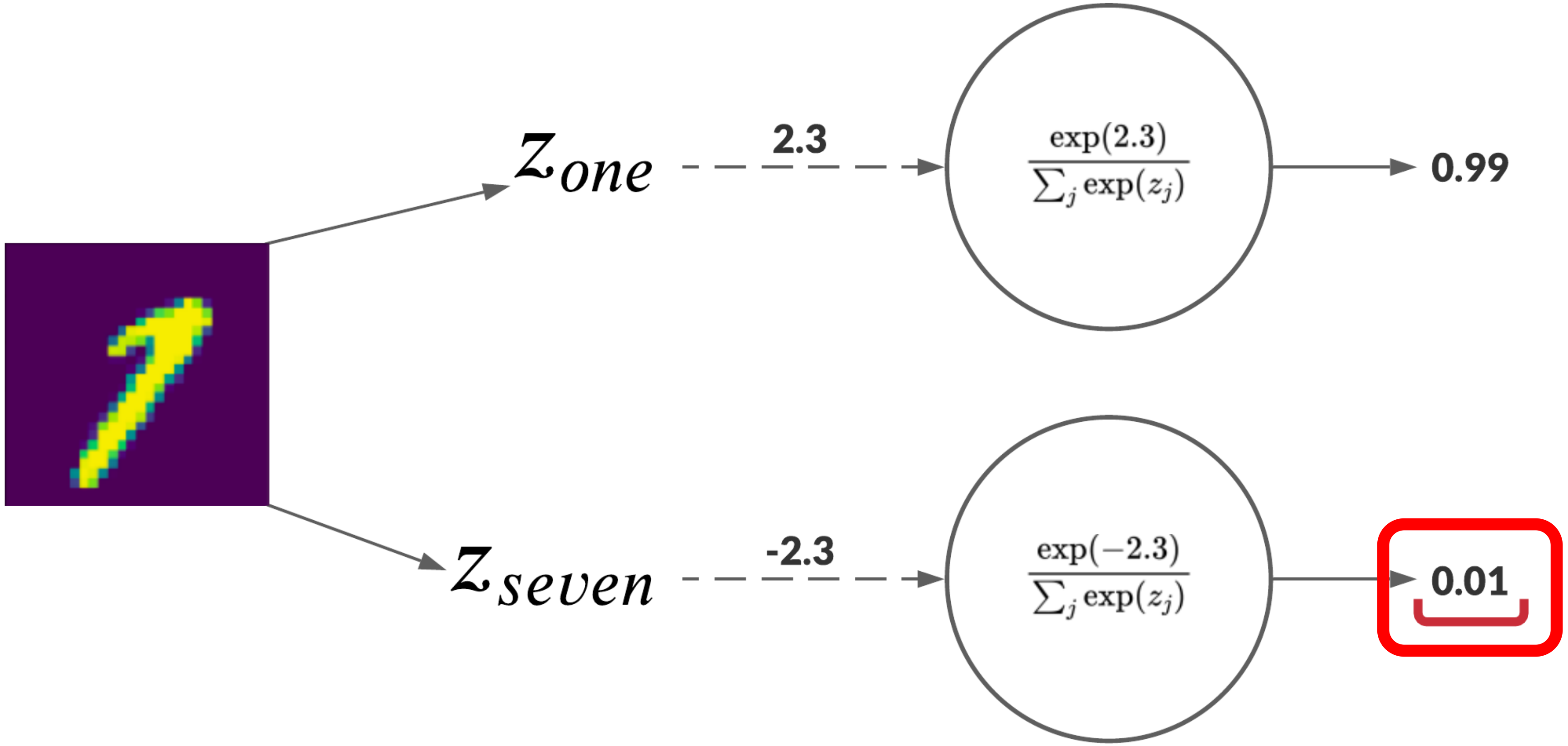
Knowledge Distillation: Teach Student the "Dark Knowledge" of Teacher



Also transfers knowledge about the relative similarity of other categories to the predicted category

Knowledge Distillation: Rebalance ("Soften") Probability Distribution Across Categories

Recall Softmax: converts scores into a probability distribution that sums to 1



Want to enhance knowledge of this relationship

https://wandb.ai/authors/knowledge-distillation/reports/Distilling-Knowledge-in-Neural-Networks--VmlldzoyMjkxODk

Knowledge Distillation: Rebalance ("Soften") Probability Distribution Across Categories

Generalized Softmax: converts scores into a probability distribution summing to 1, with temperature

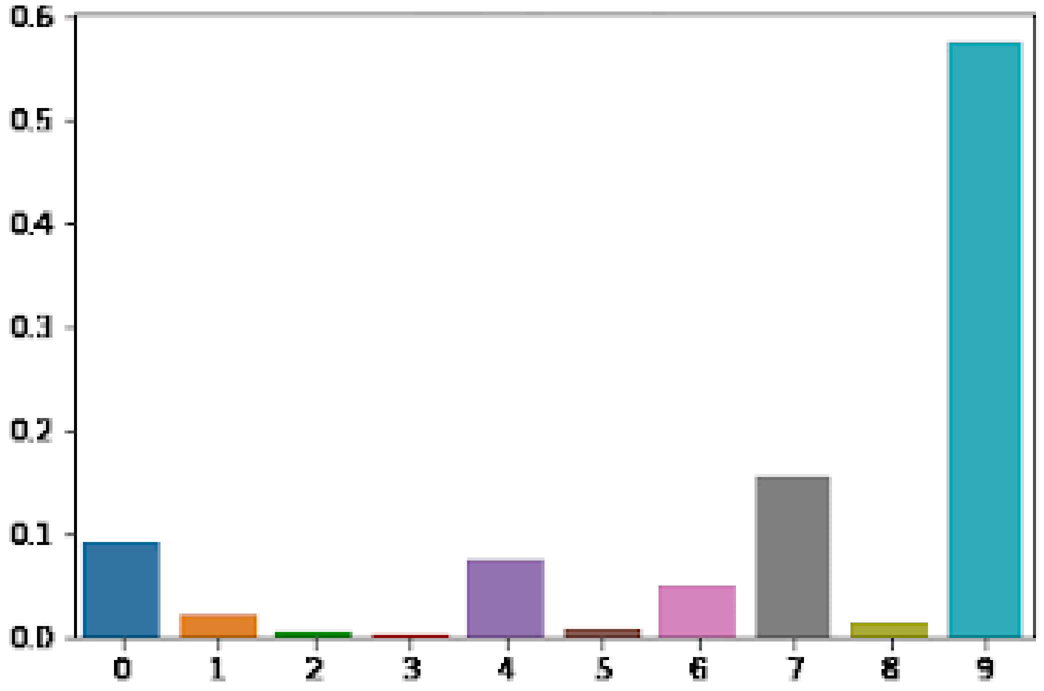
$$\sigma(\mathbf{z})_i = \frac{\exp(z_i / \boxed{T})}{\sum_j \exp(z_j / \boxed{T})}$$

What is the typical value of T used for softmax?

Idea: set the temperature to a value greater than 1

Knowledge Distillation: Rebalance (“Soften”) Probability Distribution Across Categories

Generalized Softmax: converts scores into a probability distribution summing to 1, with temperature

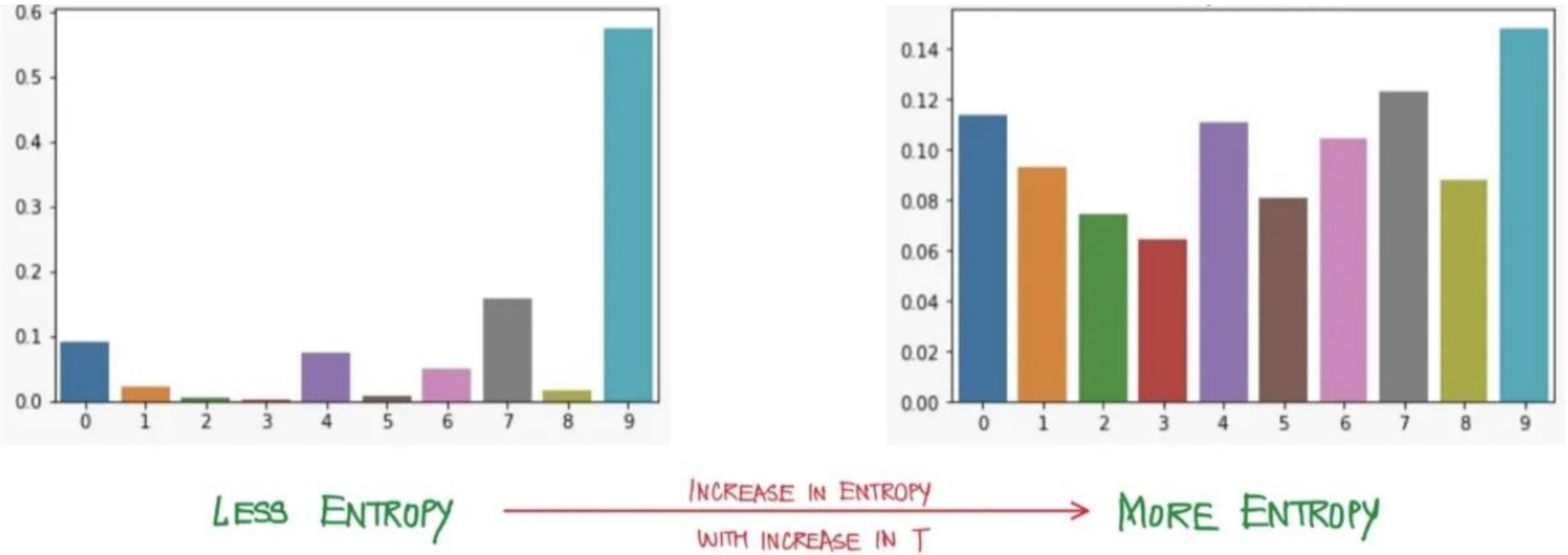


Larger T values means more information is available about which categories the teacher found similar to the predicted category

https://medium.com/@harshit158/softmax-temperature-5492e4007f71

Knowledge Distillation: Rebalance ("Soften") Probability Distribution Across Categories

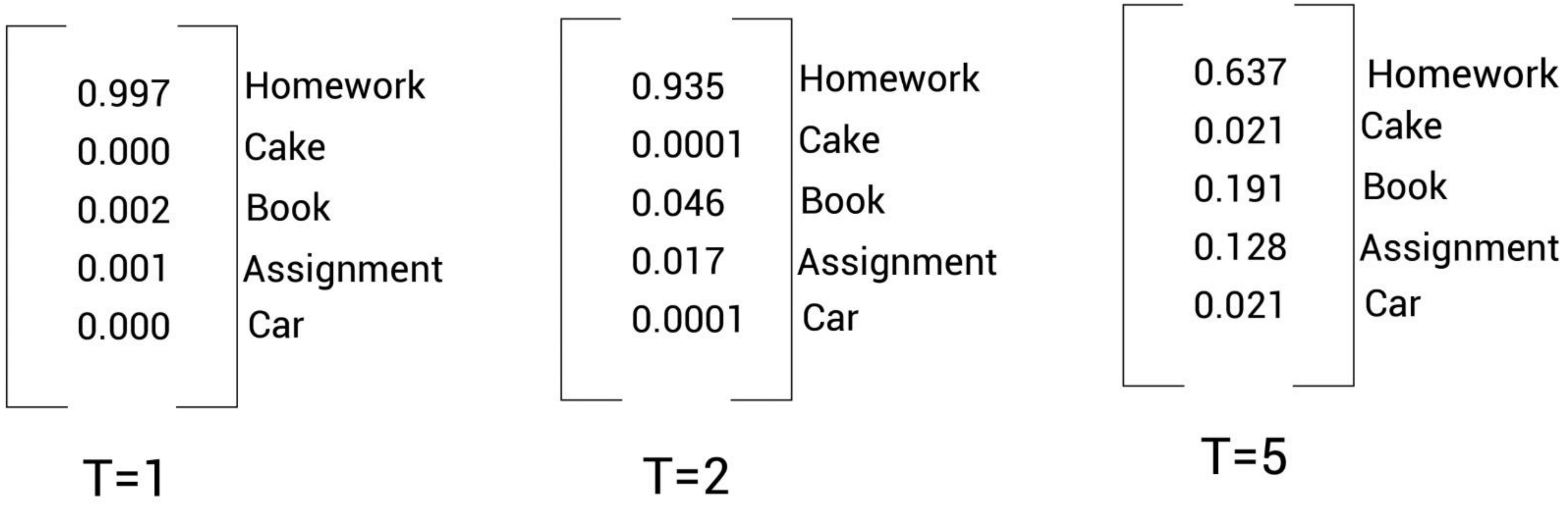
Generalized Softmax: converts scores into a probability distribution summing to 1, with temperature



https://medium.com/@harshit158/softmax-temperature-5492e4007f71

Knowledge Distillation: Rebalance ("Soften") Probability Distribution Across Categories

Generalized Softmax: converts scores into a probability distribution summing to 1, with temperature; e.g.,

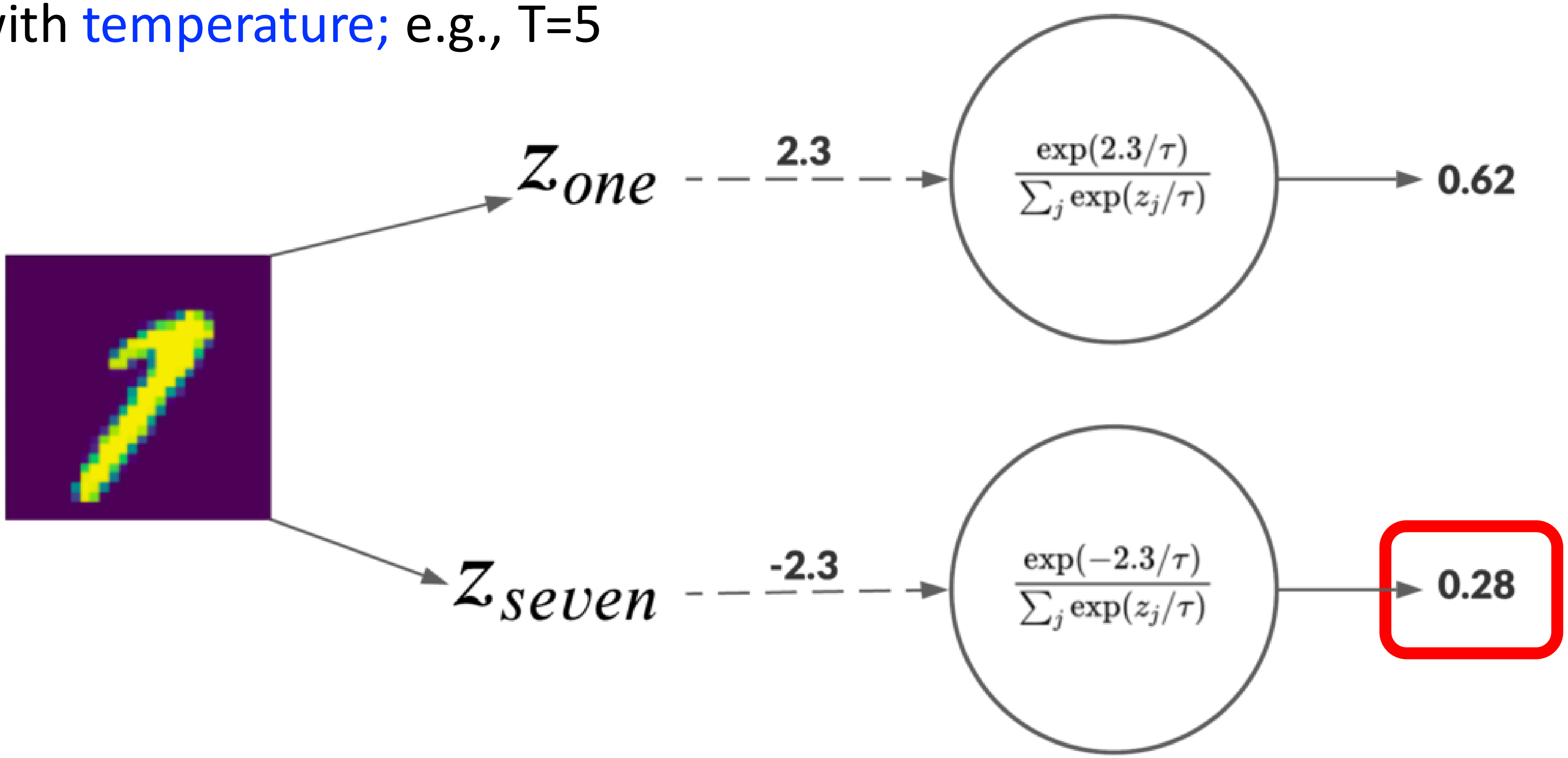


Larger T values means more information is available about which categories the teacher found similar to the predicted category

https://static.packt-cdn.com/downloads/9781838821593_ColorImages.pdf

Knowledge Distillation: Rebalance ("Soften") Probability Distribution Across Categories

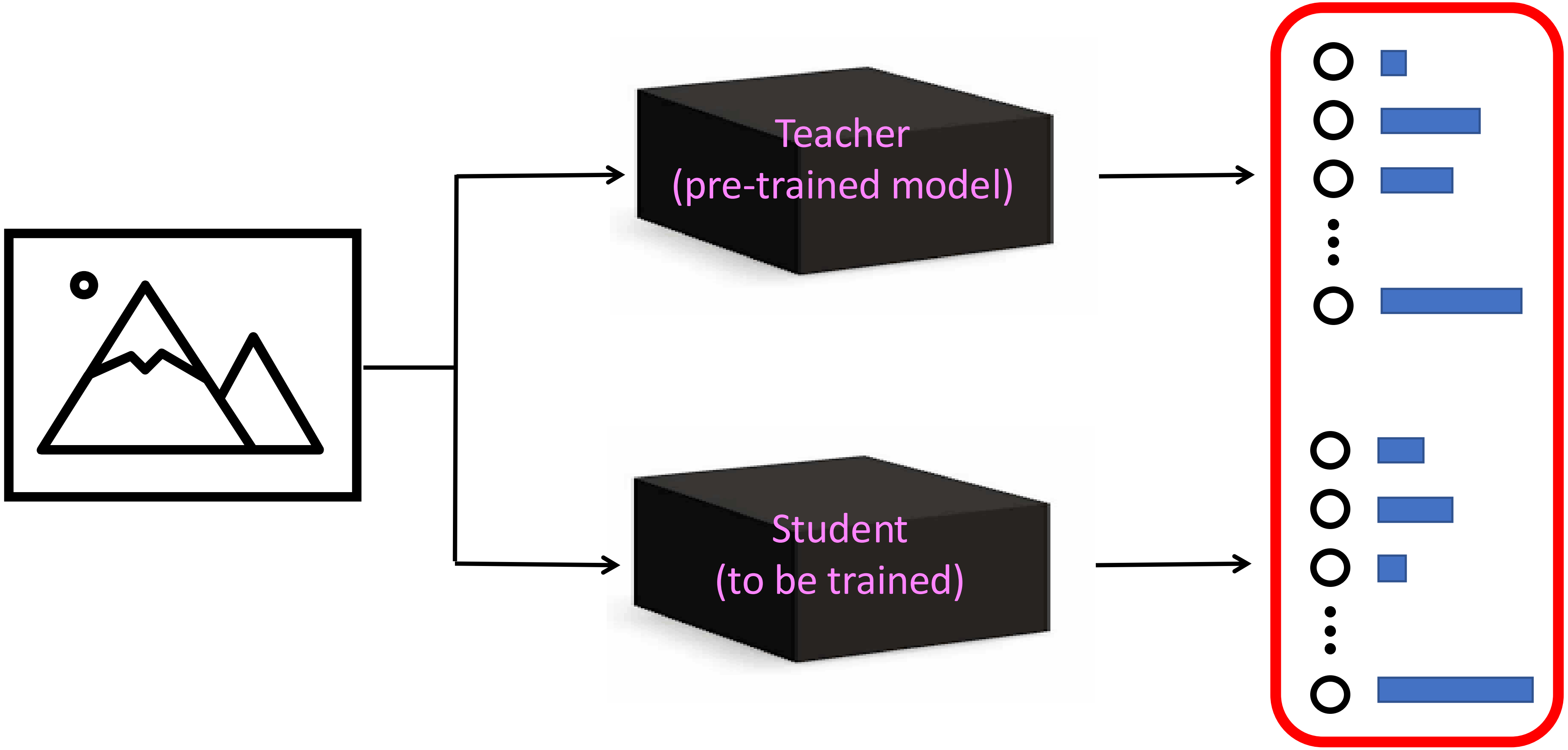
Generalized Softmax: converts scores into a probability distribution summing to 1, with temperature; e.g., T=5



Knowledge of this relationship enhanced from 0.01

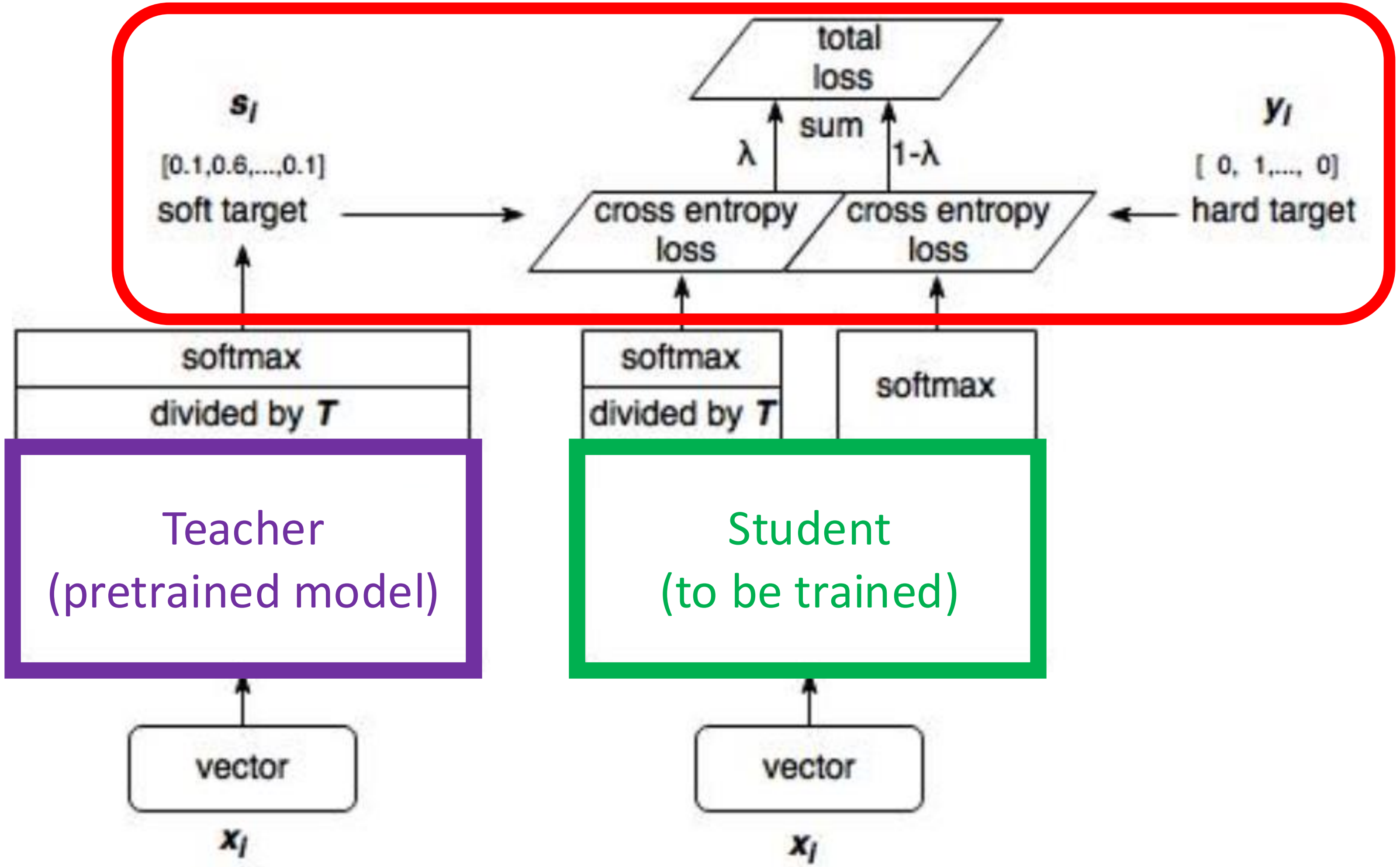
https://wandb.ai/authors/knowledge-distillation/reports/Distilling-Knowledge-in-Neural-Networks--VmlldzoyMjkxODk

Knowledge Distillation: Teach Student the "Dark Knowledge" of Teacher



"Distillation" loss computed to bring student's distribution closer to that of the teacher, using the generalized softmax equation

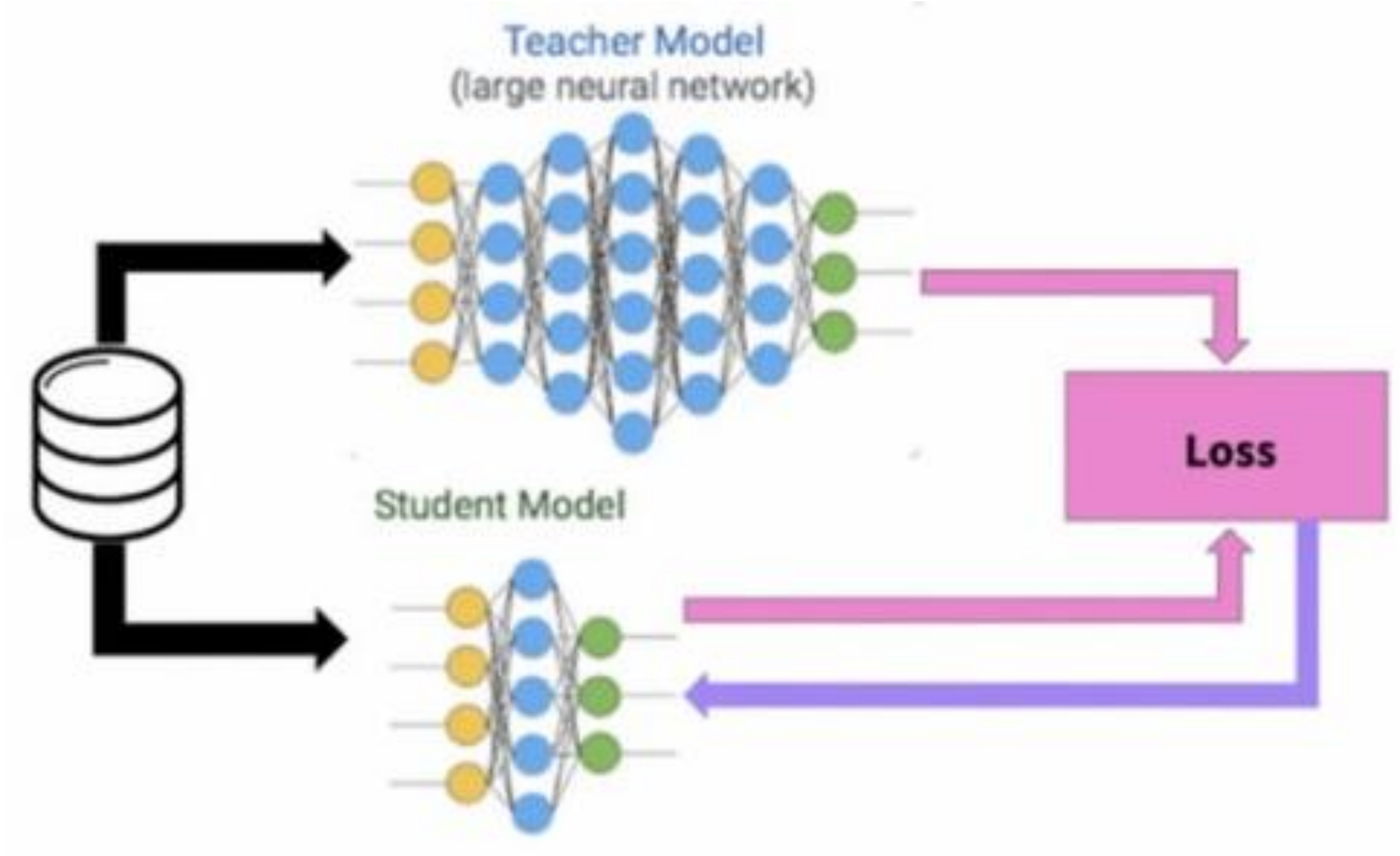
Knowledge Distillation: Teach Student the "Dark Knowledge" of Teacher



Total loss computed during training is a weighted sum of the conventional cross entropy loss and the "distillation loss"

https://blog.csdn.net/qq_22749699/article/details/79460817

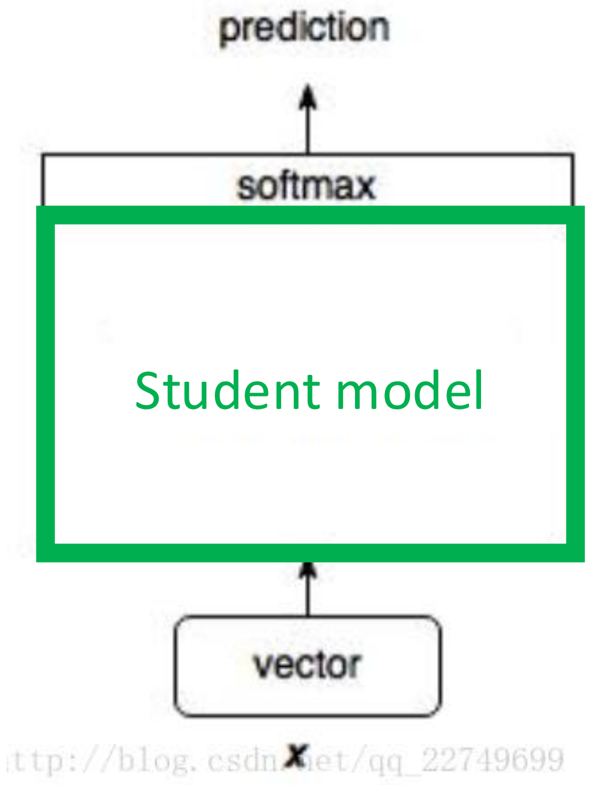
Knowledge Distillation: Teach Student the "Dark Knowledge" of Teacher



Total loss computed during training is a weighted sum of the conventional cross entropy loss and the "distillation loss"

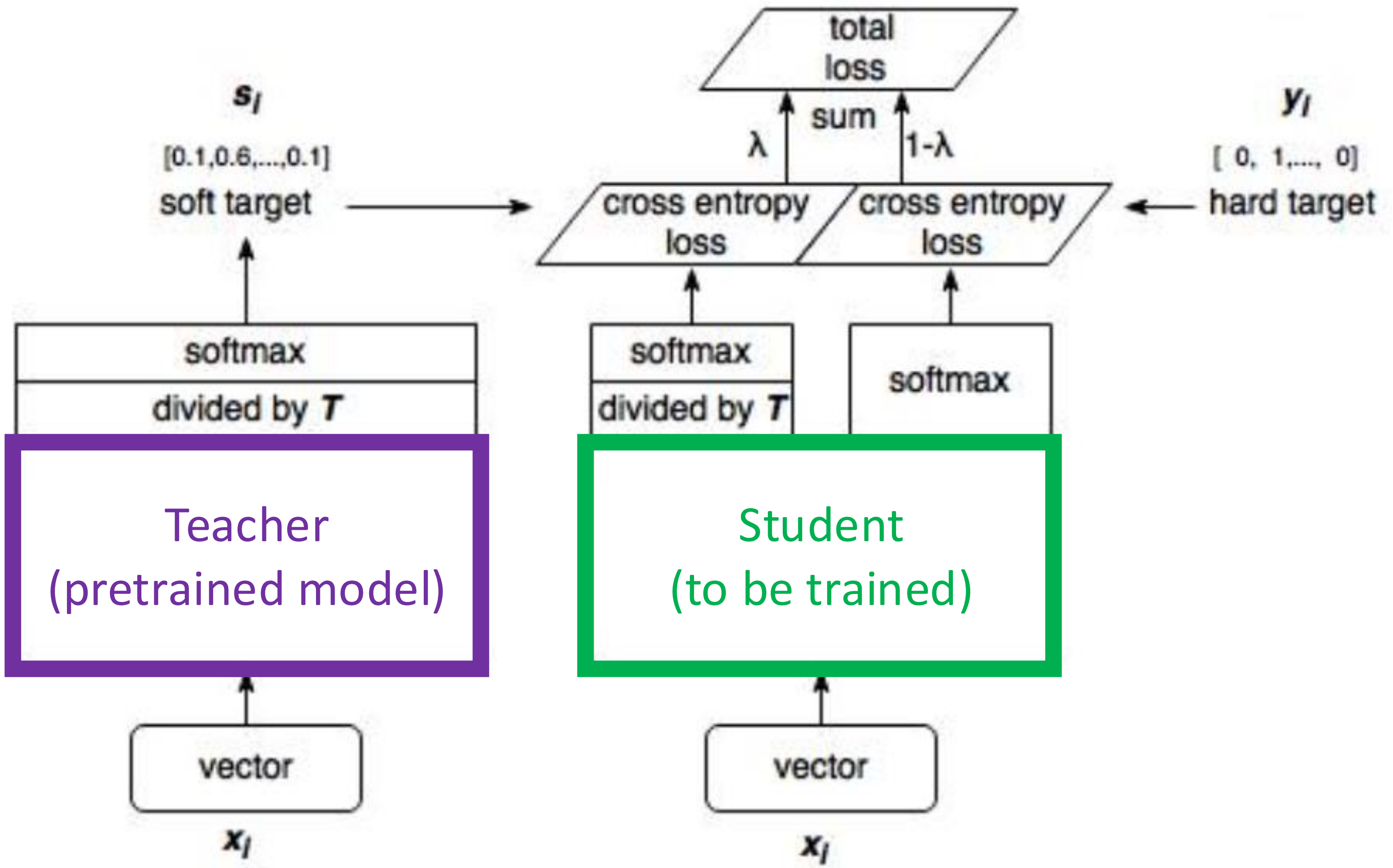
https://xailient.com/blog/4-popular-model-compression-techniques-explained/

Knowledge Distillation: At Test Time



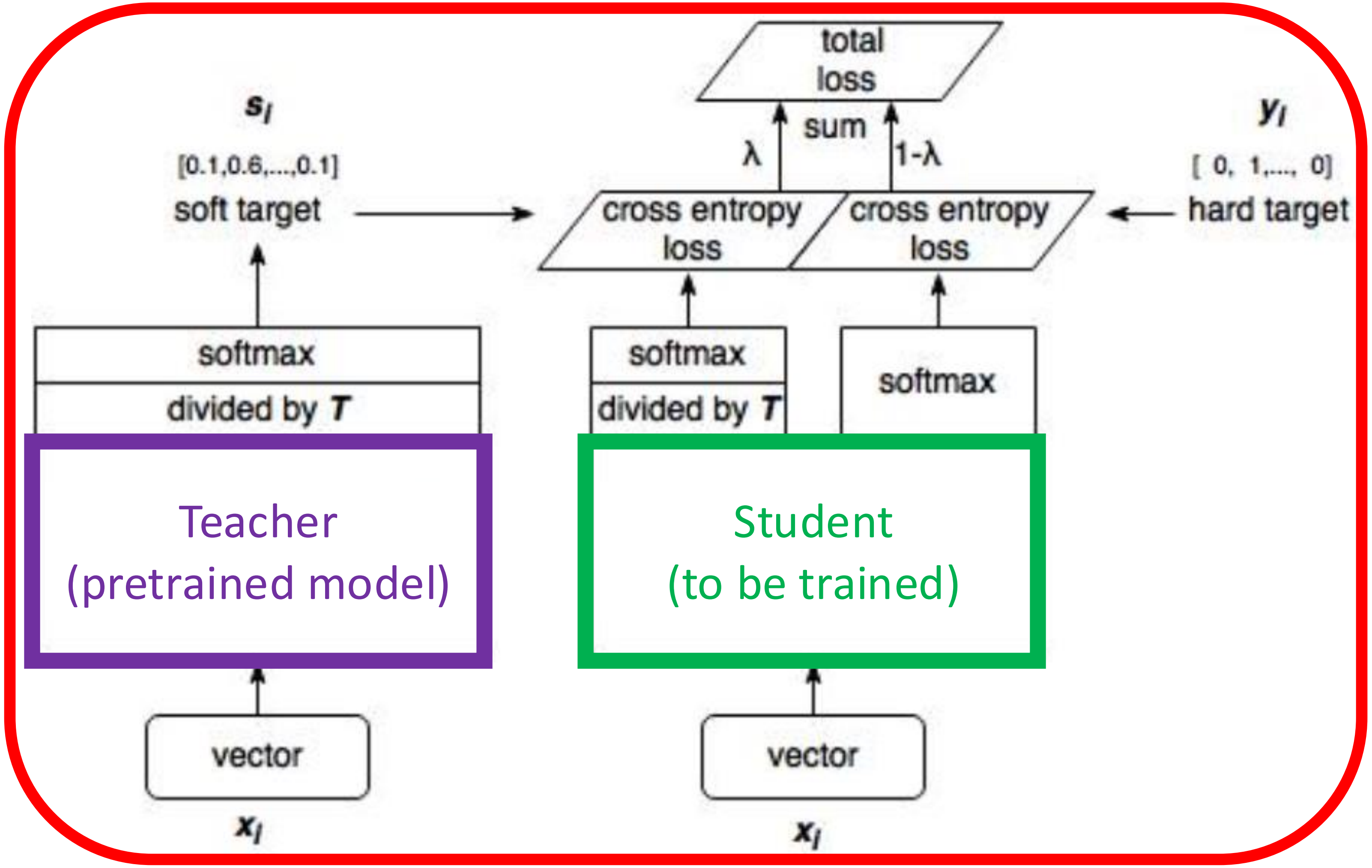
https://blog.csdn.net/qq_22749699/article/details/79460817

Arguably, Any Neural Network Student Could Learn from Any Neural Network Teacher



https://blog.csdn.net/qq_22749699/article/details/79460817

Arguably, Any Neural Network Student Could Learn from Any Neural Network Teacher

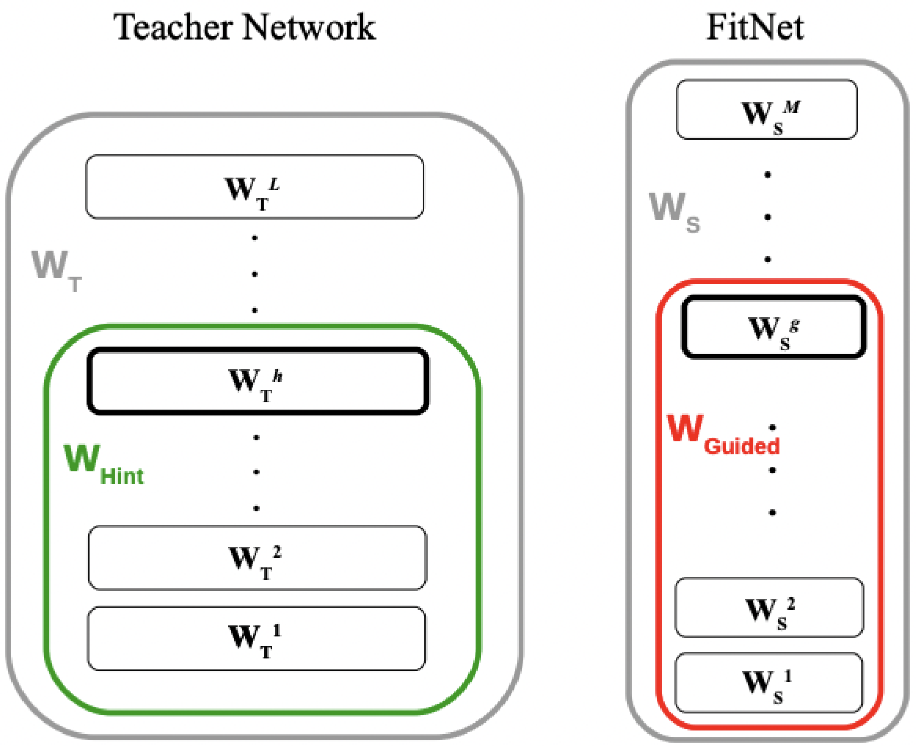


Knowledge distillation is a type of transfer learning

https://blog.csdn.net/qq_22749699/article/details/79460817

Knowledge Distillation Enhancement: Hints

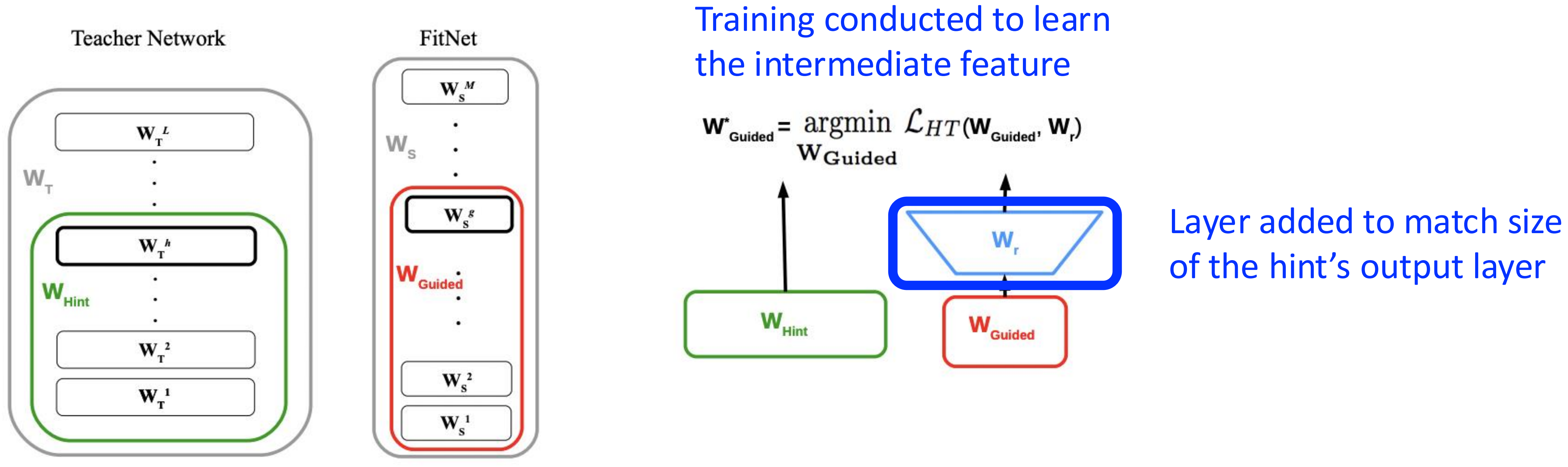
Encourage student (FitNet) to mimic the teacher's feature responses; e.g., output of guided layer should match the output of hint layer



Romero et al. Fitnets: Hints for thin deep nets. ICLR 2015.

Knowledge Distillation Enhancement: Hints

Encourage student (FitNet) to mimic the teacher's feature responses; e.g., output of guided layer should match the output of hint layer



(a) Teacher and Student Networks

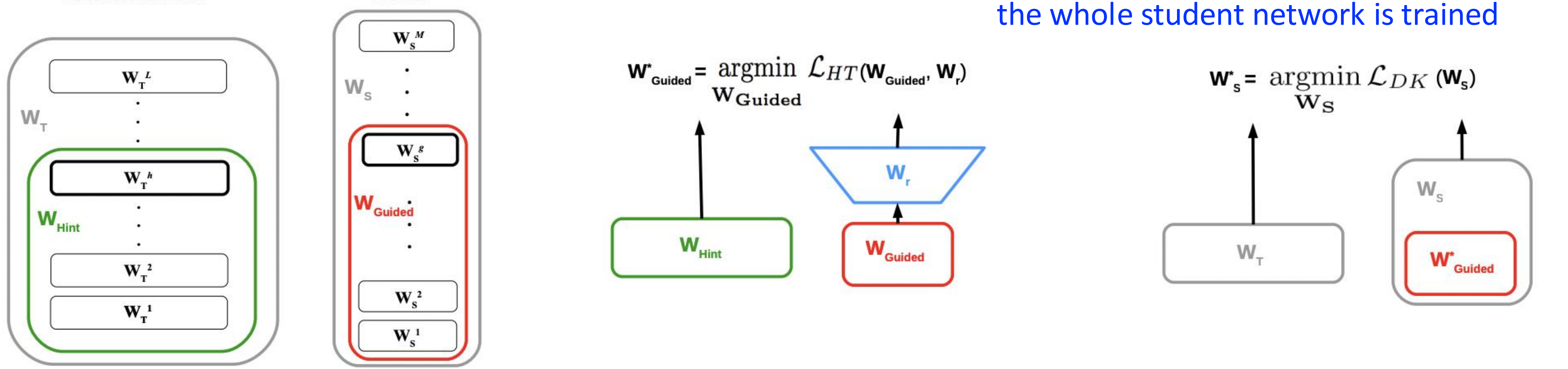
(b) Hints Training

Romero et al. Fitnets: Hints for thin deep nets. ICLR 2015.

Knowledge Distillation Enhancement: Hints

Encourage student (FitNet) to mimic the teacher's feature responses; e.g., output of guided layer should match the output of hint layer

After learning the intermediate features, the whole student network is trained



(a) Teacher and Student Networks (b) Hints Training (c) Knowledge Distillation

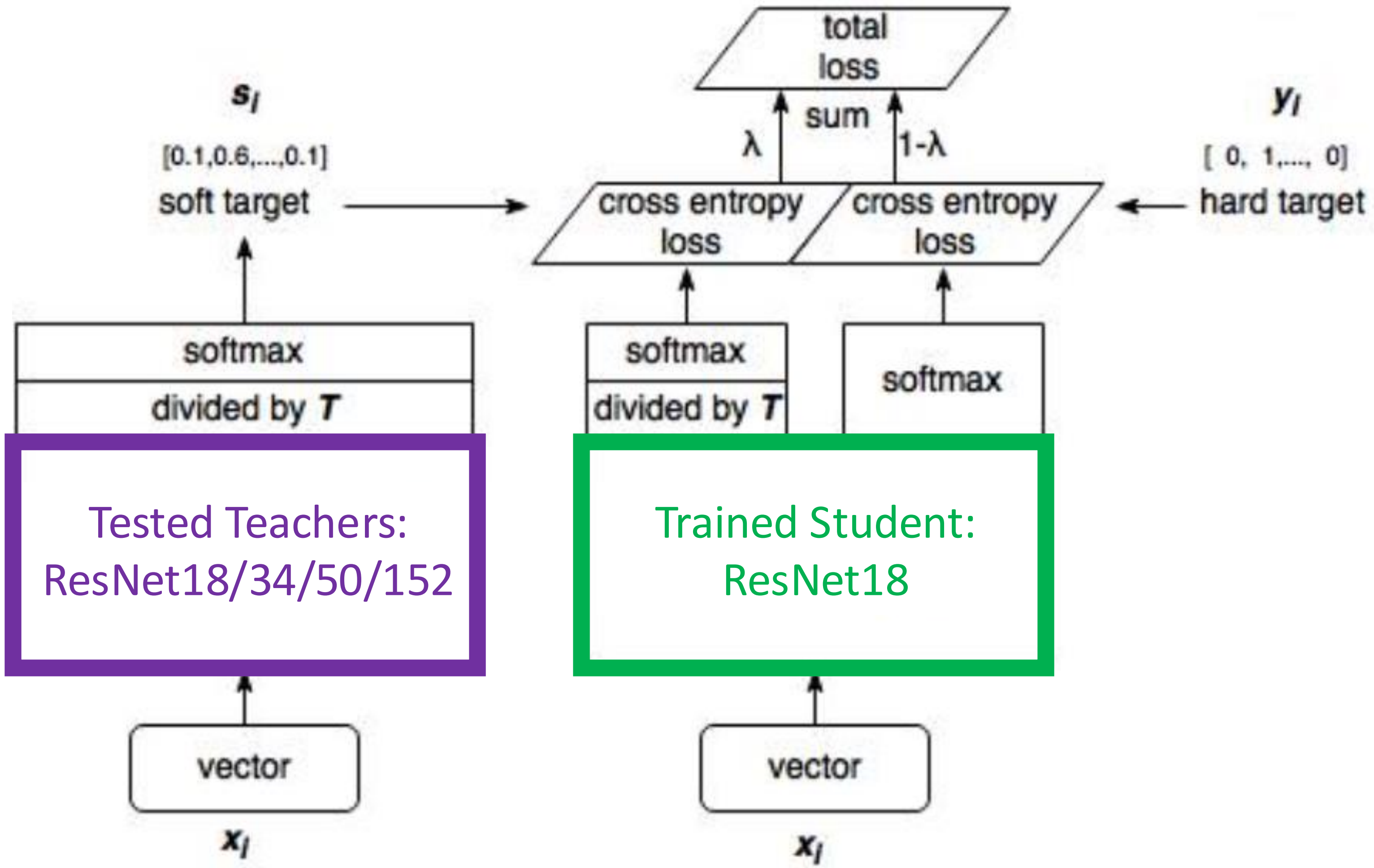
Romero et al. Fitnets: Hints for thin deep nets. ICLR 2015.

Example: Predict Category from 1000 Options

- Evaluation metric: % correct (top-1 and top-5 predictions)
- Dataset: ~1.5 million images
- Source: images scraped from search engines, such as Flickr, and labeled by crowdworkers

J. Deng, W. Dong, R. Socher, L. Li, K. Li and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. 2009

Example: Do Bigger, More Accurate Models Make Better Teachers?



Cho and Hariharan. On the Efficacy of Knowledge Distillation. ICCV 2019
Figure source: https://blog.csdn.net/qq_22749699/article/details/79460817

Example: Do Bigger, More Accurate Models Make Better Teachers?

(% = Top-1 error rates)

Teacher	Teacher Error (%)	Student Error (%)
ResNet18	30.24	30.57
ResNet34	26.70	30.79
ResNet50	23.85	30.95

What is the student's performance trend from larger, more accurate teachers?

Cho and Hariharan. On the Efficacy of Knowledge Distillation. ICCV 2019

Example: Do Bigger, More Accurate Models Make Better Teachers?

(% = Top-1 error rates)

Teacher	Teacher Error (%)	Student Error (%)
-	-	30.24
ResNet18	30.24	30.57
ResNet34	26.70	30.79
ResNet50	23.85	30.95

Student performance not only drops for larger teachers but the models distilled from teachers perform worse than training the student from scratch!

Cho and Hariharan. On the Efficacy of Knowledge Distillation. ICCV 2019

Example: Why Might Student Performance Drop as Teacher Size Grows?

1. More accurate models are more confident and so need higher temperatures to learn the "dark knowledge" of category relationships

2. Student mimics teacher but the loss function is mismatched from the evaluation metric

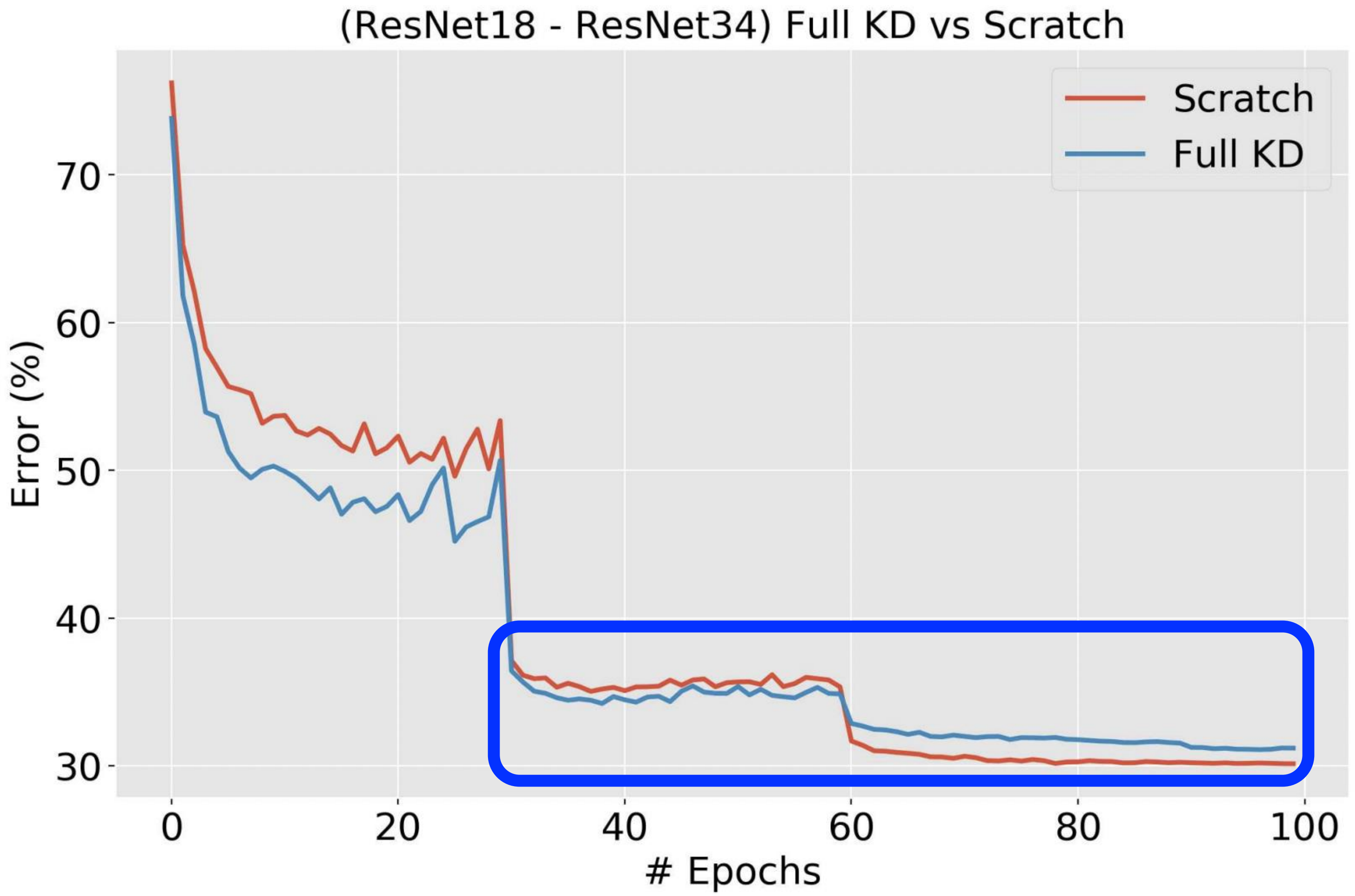
3. Student fails to accurately mimic teacher

Experimental analysis suggests this is the reason

Cho and Hariharan. On the Efficacy of Knowledge Distillation. ICCV 2019

Example: Why Might Students Fail to Mimic Teachers?

Hypothesis: student is underfitting from smaller capacity and so "minimizing one loss (KD loss) at the expense of the other (cross entropy loss)"

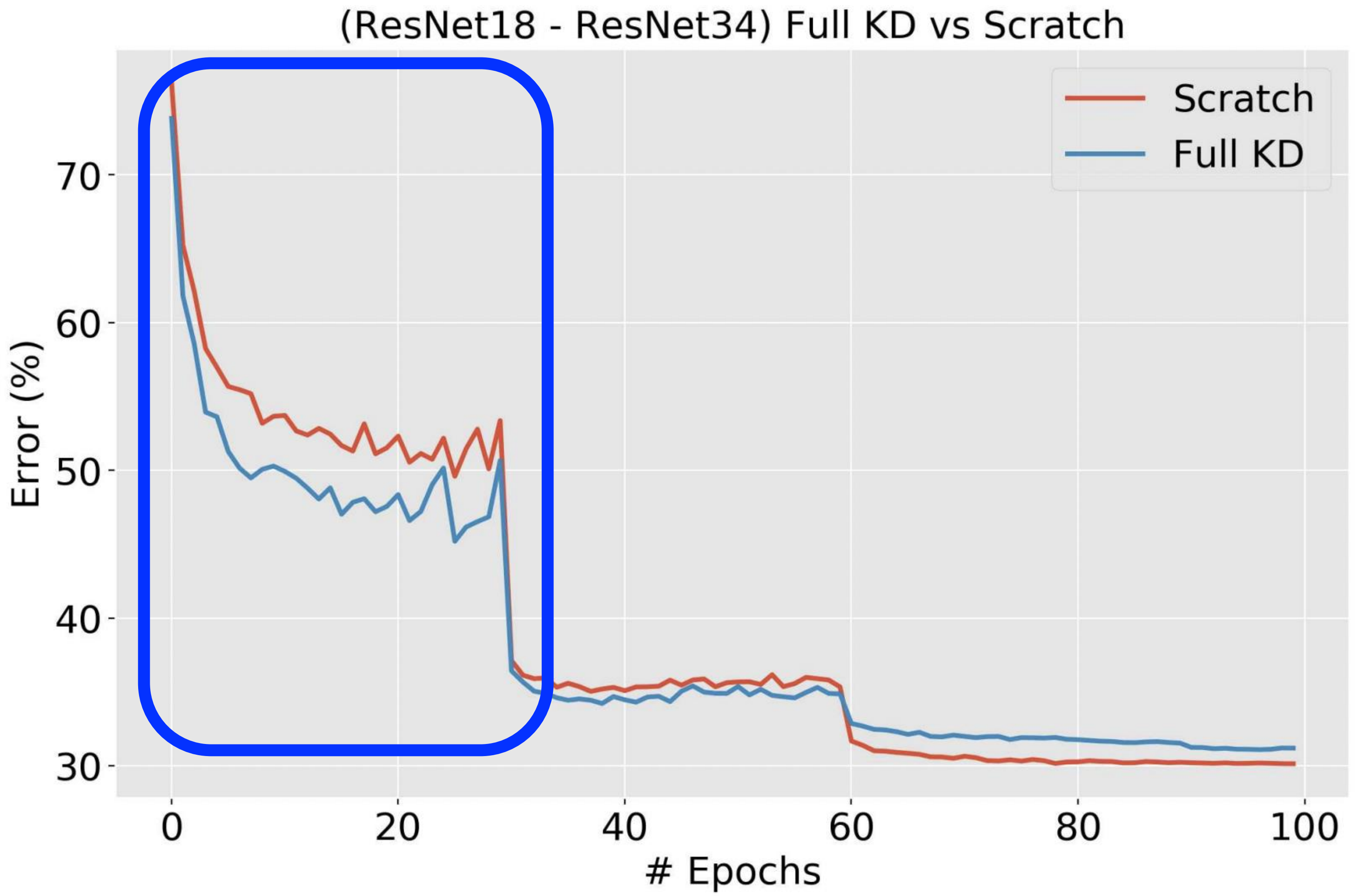


Cho and Hariharan. On the Efficacy of Knowledge Distillation. ICCV 2019

Example: Why Might Students Fail to Mimic Teachers?

How to overcome this issue?

- Early stopping with KD loss (ESKD) to leverage its benefit at the start of training



Cho and Hariharan. On the Efficacy of Knowledge Distillation. ICCV 2019

Example: How Does ESKD Compare To Training A Student from Scratch?

Teacher	Top-1 Error (%, Test)
ResNet18	30.57
ResNet18 (ES KD)	29.01
ResNet34	30.79
ResNet34 (ES KD)	29.16
ResNet50	30.95
ResNet50 (ES KD)	29.35

Training a model with early stopping knowledge distillation loss leads to better results than training from scratch!

Cho and Hariharan. On the Efficacy of Knowledge Distillation. ICCV 2019

Example: Are Results from ESKD Better When Using Bigger, More Accurate Models As Teachers?

Teacher	Top-1 Error (%, Test)
ResNet18	30.57
ResNet18 (ES KD)	29.01
ResNet34	30.79
ResNet34 (ES KD)	29.16
ResNet50	30.95
ResNet50 (ES KD)	29.35

No; the student may still be struggling with underfitting due to an insufficient representational capacity

Cho and Hariharan. On the Efficacy of Knowledge Distillation. ICCV 2019

Example: To Address The Capacity Problem Why Not Instead Distill to Intermediate Sizes?

Performs almost identically to a model that is distilled directly from a large to small size; does not address the core problem:

The student must be in the solution space of the teacher

Cho and Hariharan. On the Efficacy of Knowledge Distillation. ICCV 2019

What's Currently Interesting? e.g.,

What Knowledge Gets Distilled in Knowledge Distillation?

Utkarsh Ojha* **Yuheng Li*** **Anirudh Sundara Rajan***

Yingyu Liang **Yong Jae Lee**

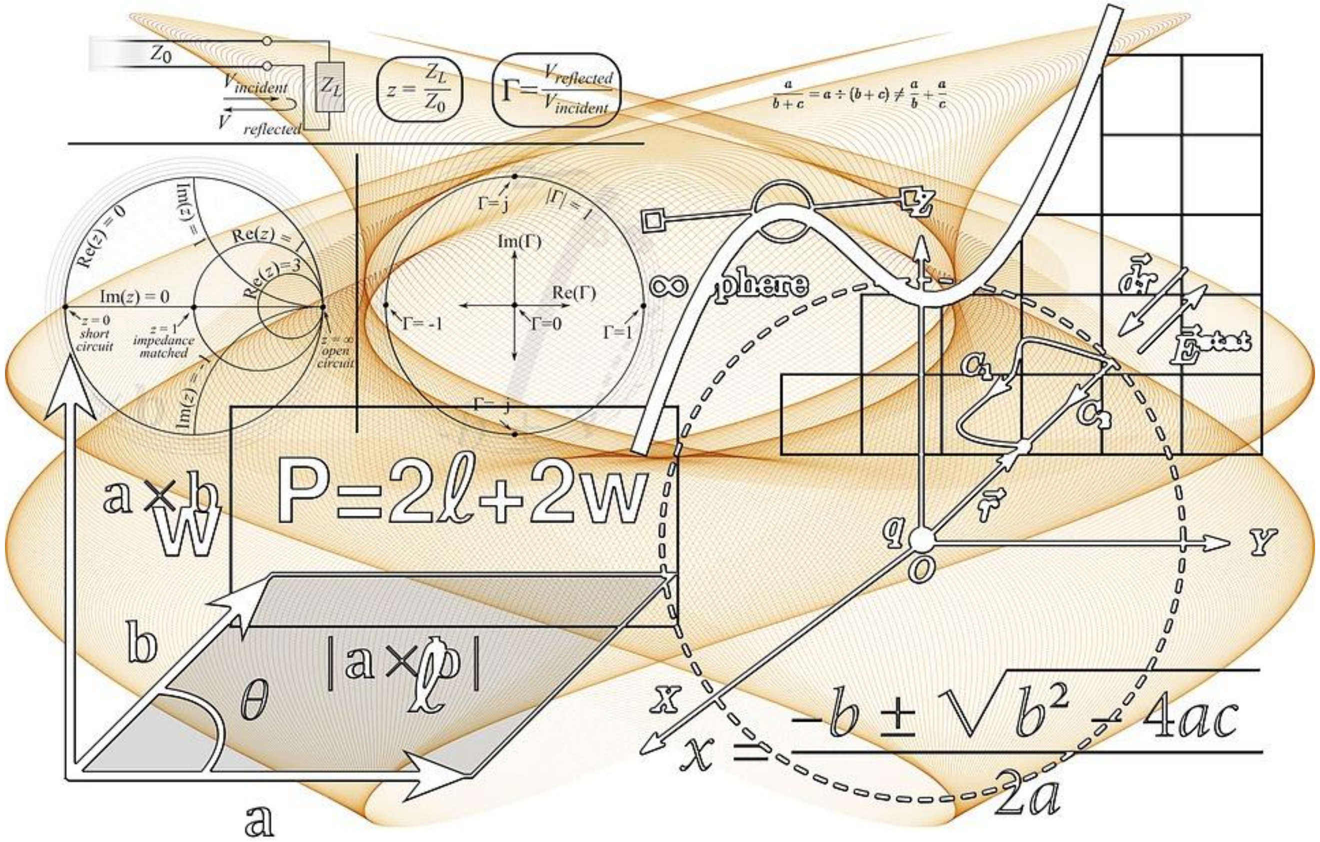
(Neurips 2024)

Efficient Computer Vision

- Motivation
- Model Compression
- Curriculum Learning
- Active Learning
- Faculty Course Questionnaire (FCQ)

Intuition: How to Teach a Child Math?

Random Order of Examples



Meaningful Order of Examples

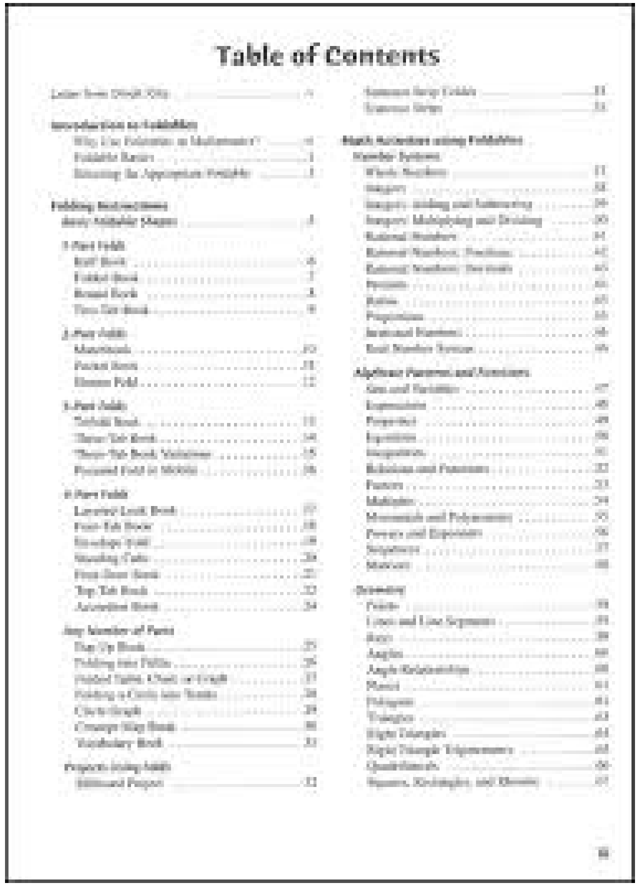


Table of Contents

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

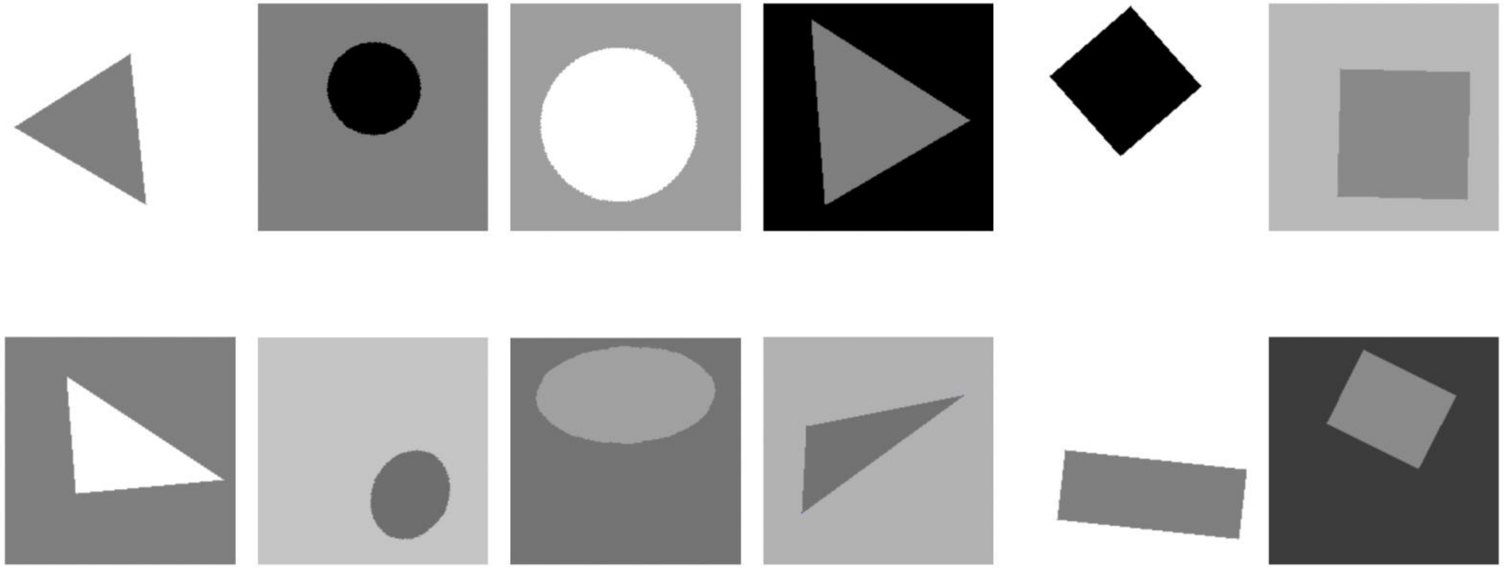
Jeffrey L. Elman. Learning and development in neural networks: The importance of starting small. Cognition, 1993.

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

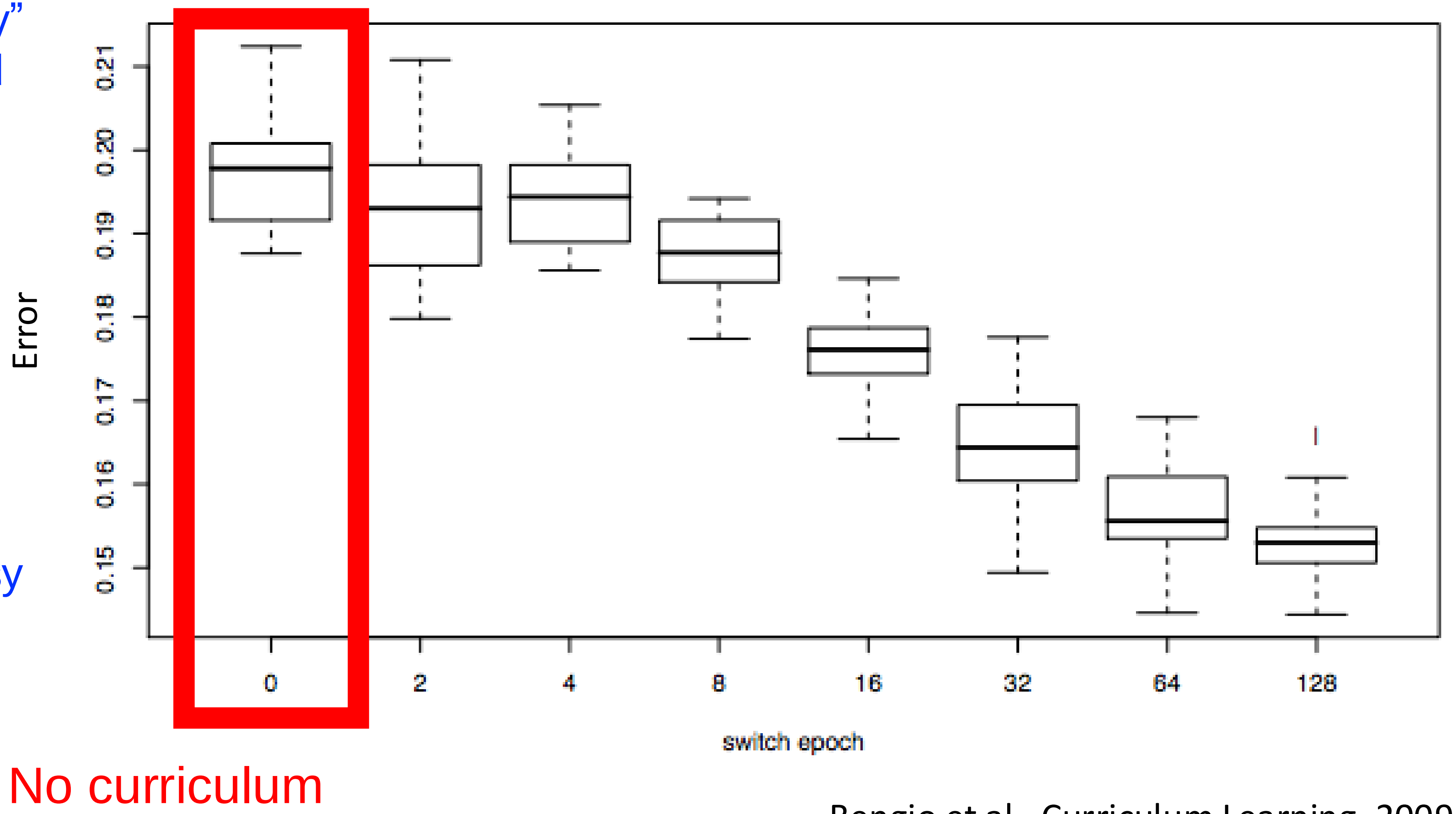
Bengio et al., Curriculum Learning, 2009

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

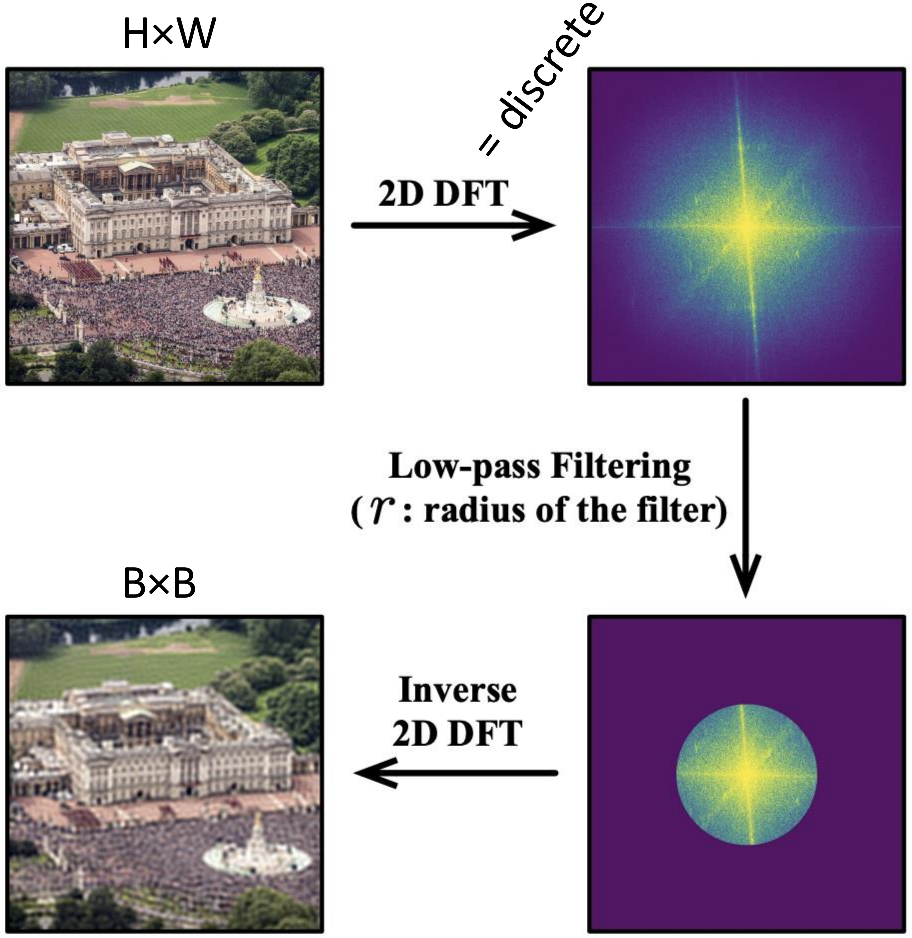
Yulin Wang[1*] Yang Yue[1*] Rui Lu[1] Tianjiao Liu[2] Zhao Zhong[2]
Shiji Song[1] Gao Huang[1,3✉]
[1]Department of Automation, BNRist, Tsinghua University [2]Huawei Technologies Ltd. [3]BAAI
{wang-yl19, yueyang22}@mails.tsinghua.edu.cn, gaohuang@tsinghua.edu.cn

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

= discrete Fourier transform



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

B×B patch cropped in frequency domain

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?

Efficient Computer Vision

- Motivation
- Model Compression
- Curriculum Learning
- Active Learning
- Faculty Course Questionnaire (FCQ)

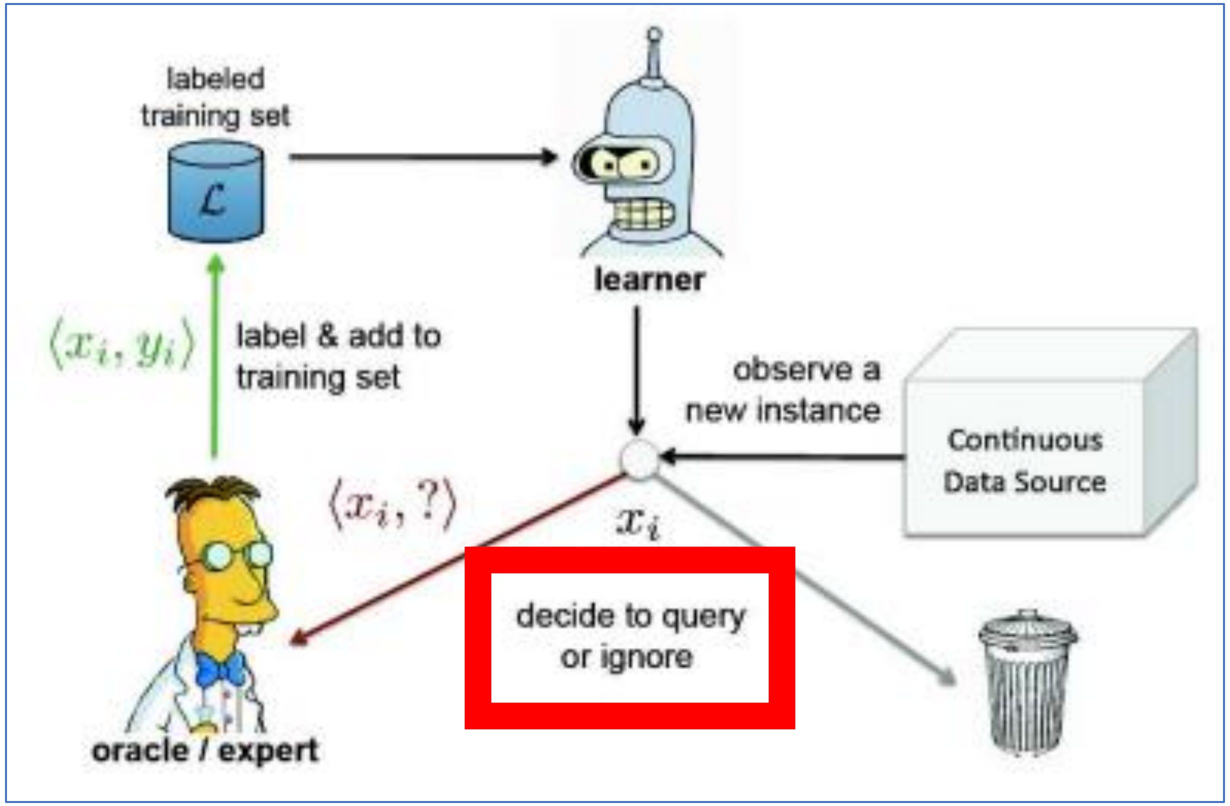
How to teach machines with minimal human supervision?

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

e.g., limited funding

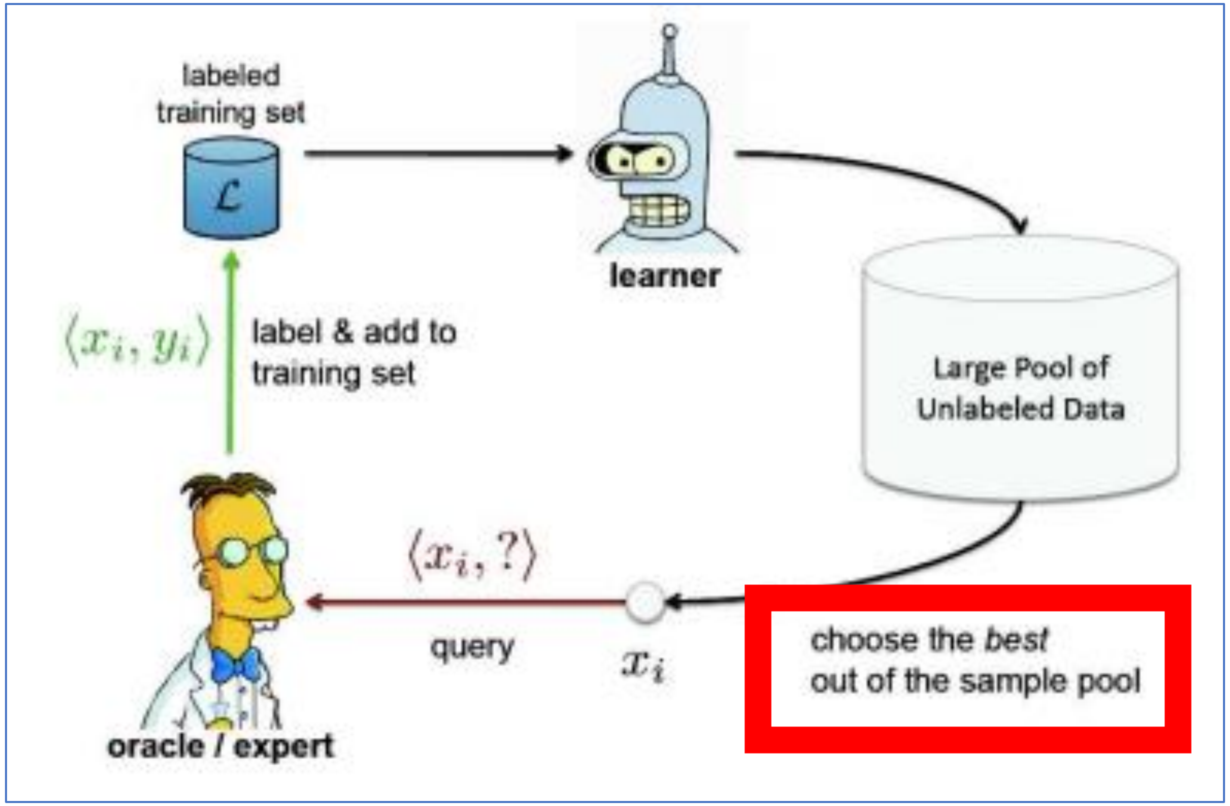
Idea: Choose Most Informative Data to Label

Stream-Based



Consider one example at a time

Pool-Based



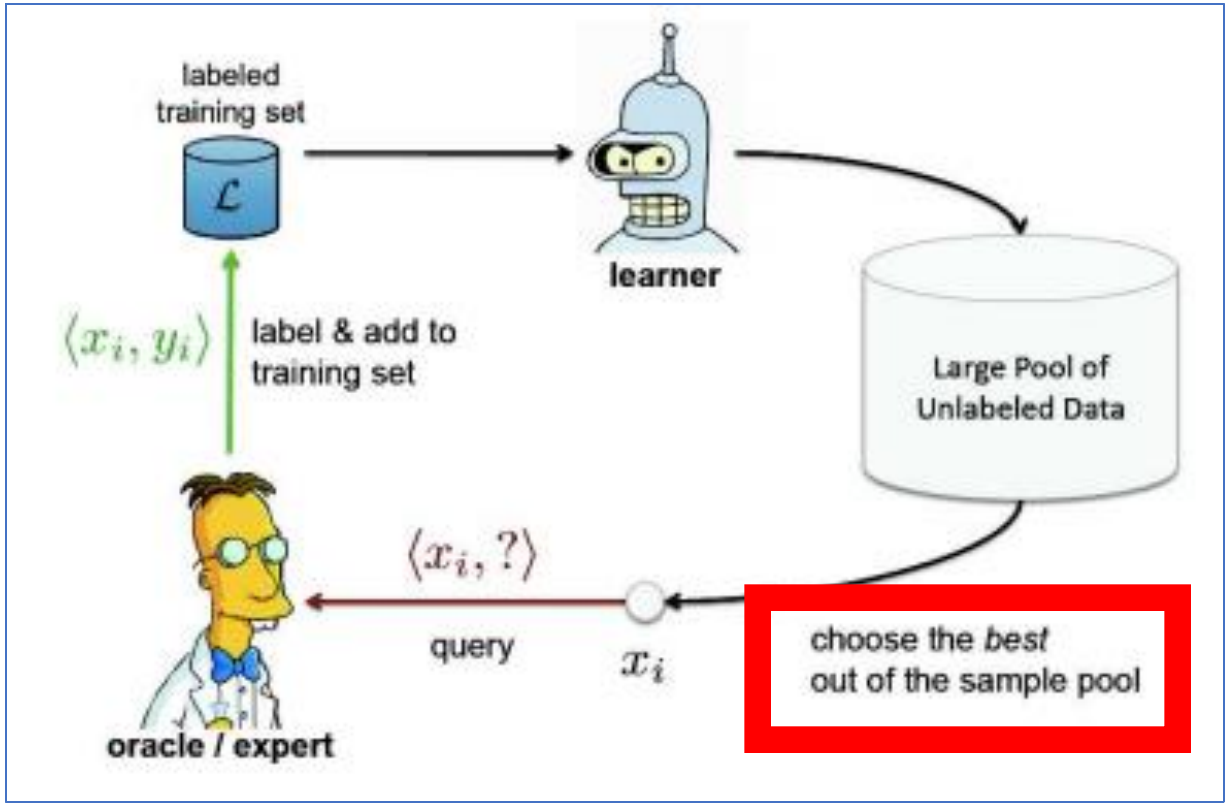
Consider many examples at a time

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

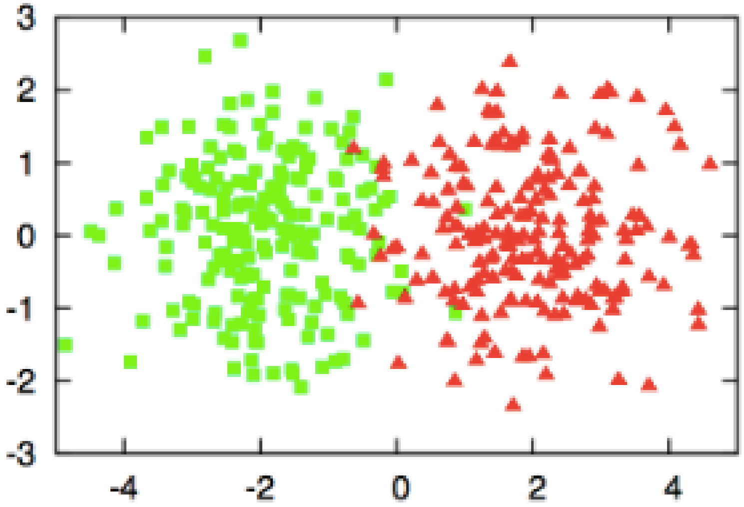
Image Credit: https://www.cs.utah.edu/~piyush/teaching/10-11-slides.pdf

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

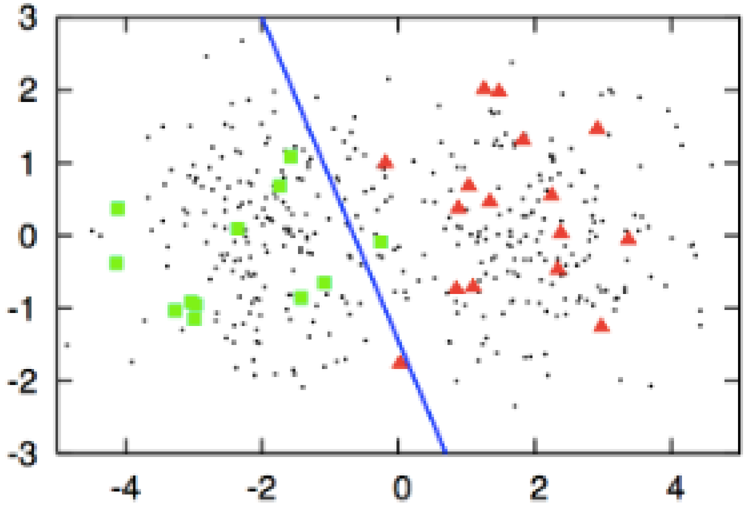
Common Approach: Uncertainty Sampling

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

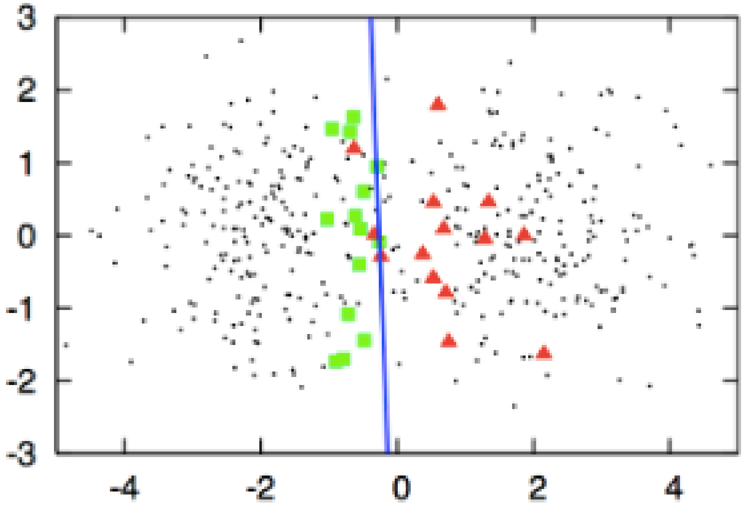
True Representation (Assume Labels Are Not Known)



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.,

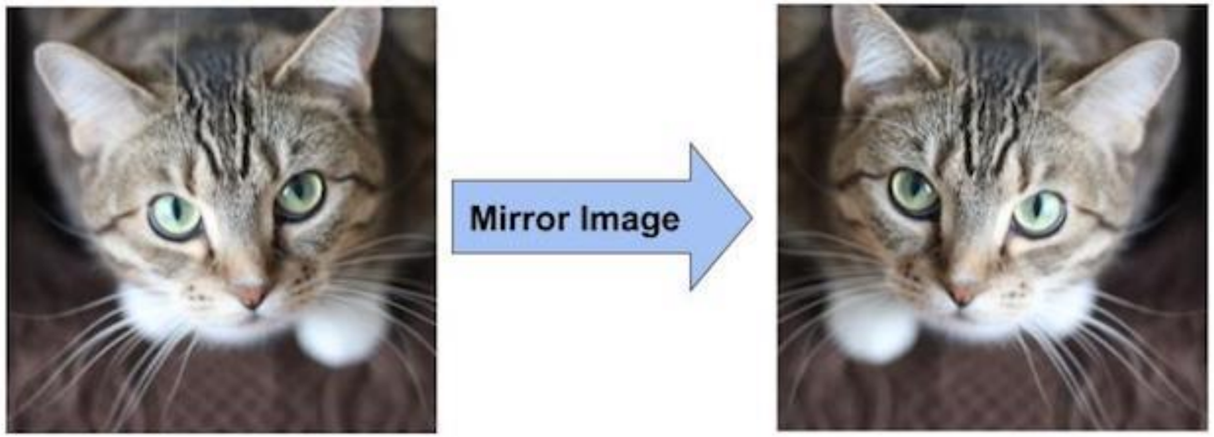
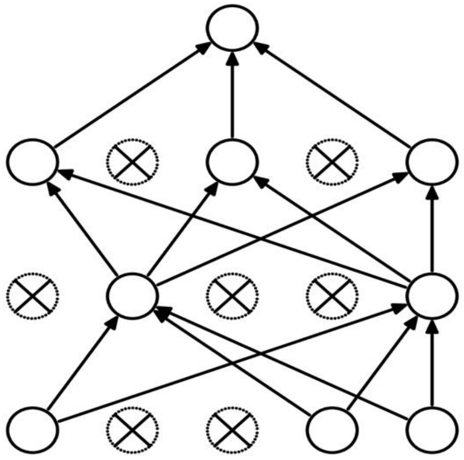


Figure Source: https://learnopencv.com/understanding-alexnet/

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

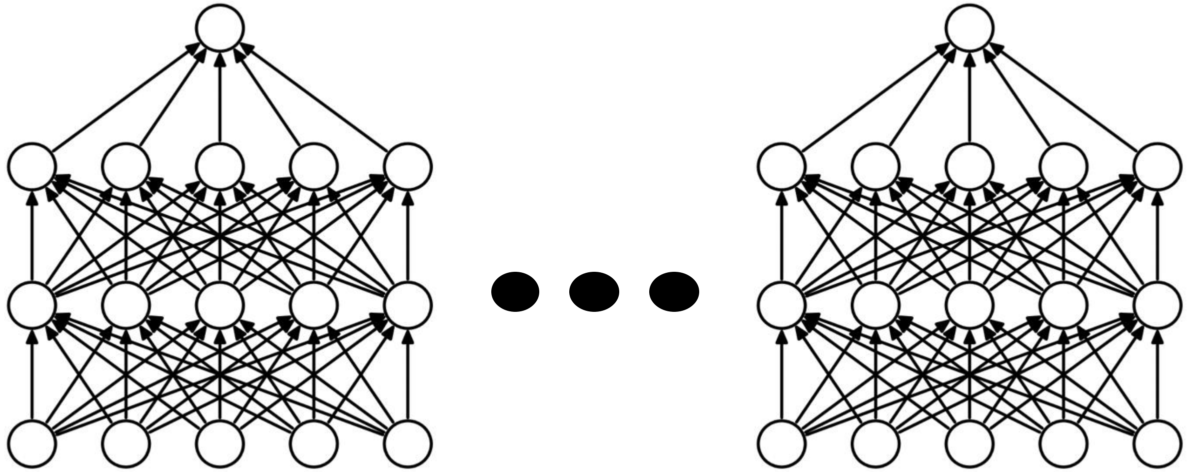


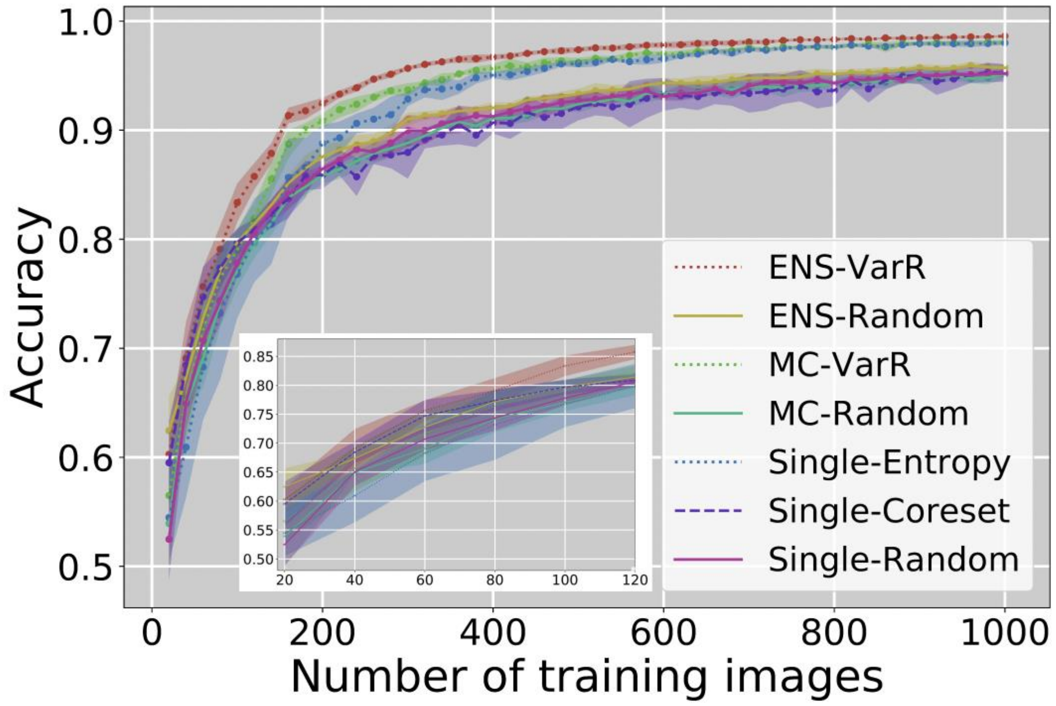
Figure Source: Srivastava et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research. 2014

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

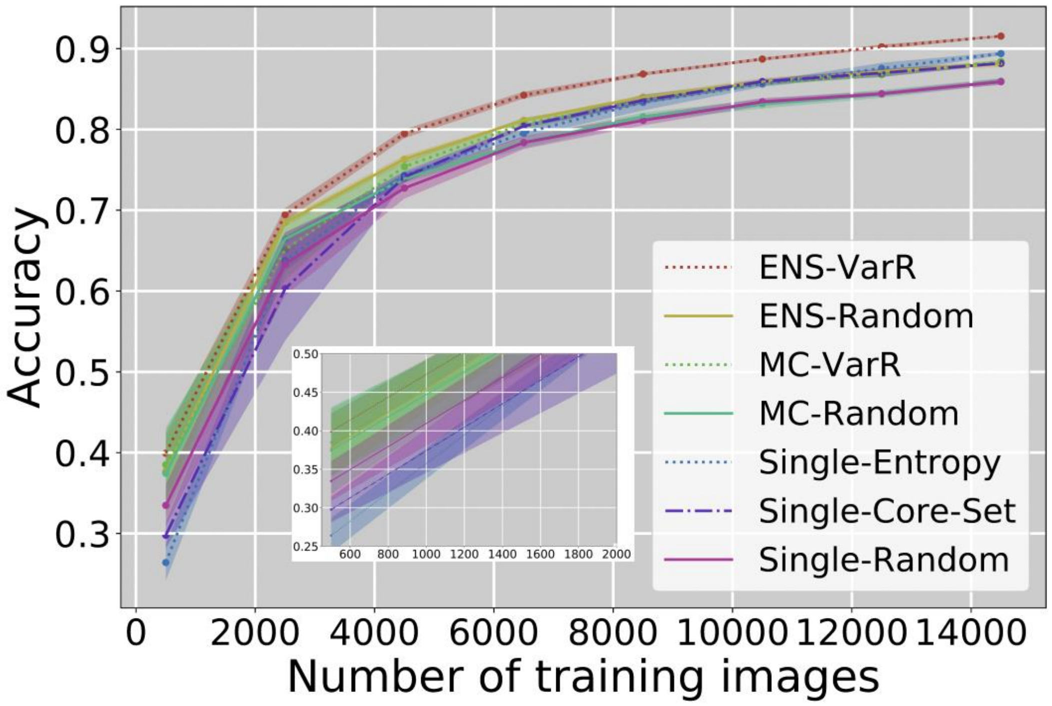
Beluch et al. The power of ensembles for active learning in image classification. CVPR 2018

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

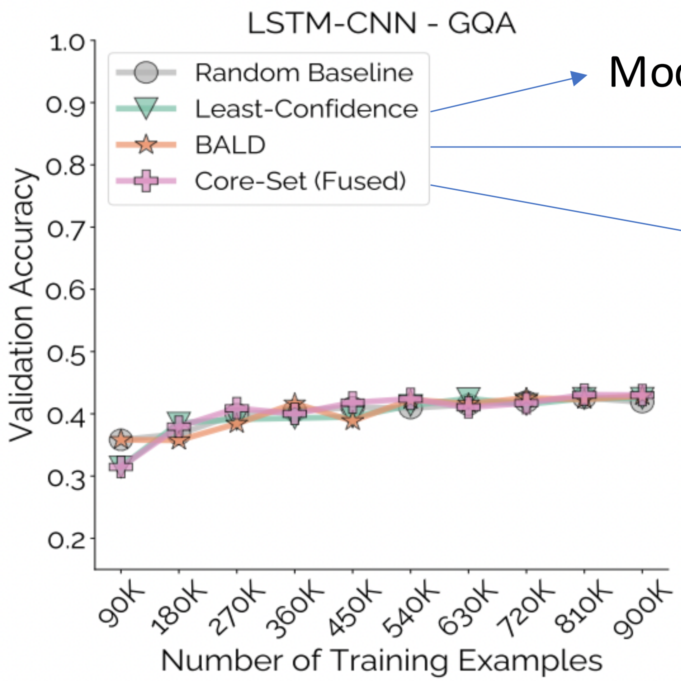


(b) CIFAR-10 on DenseNet

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.,

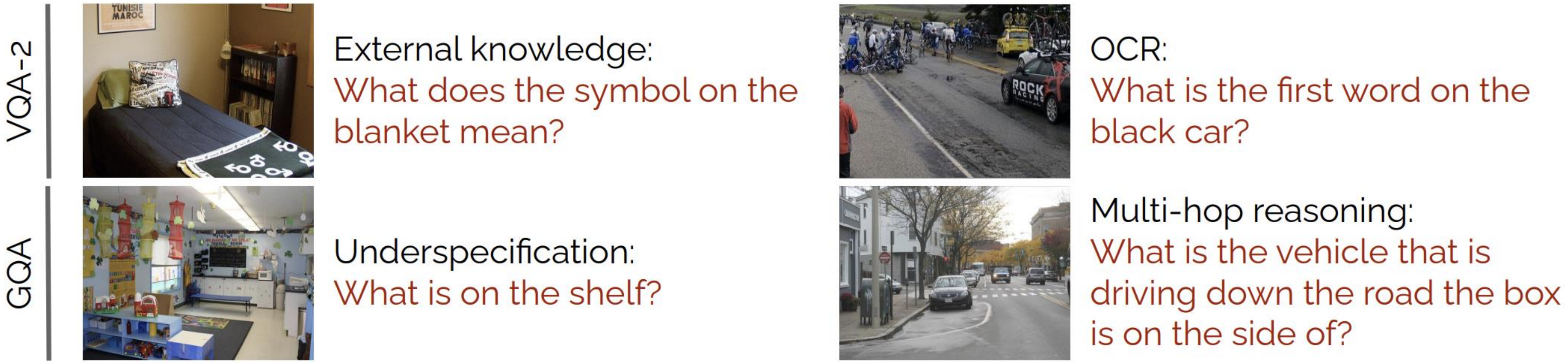
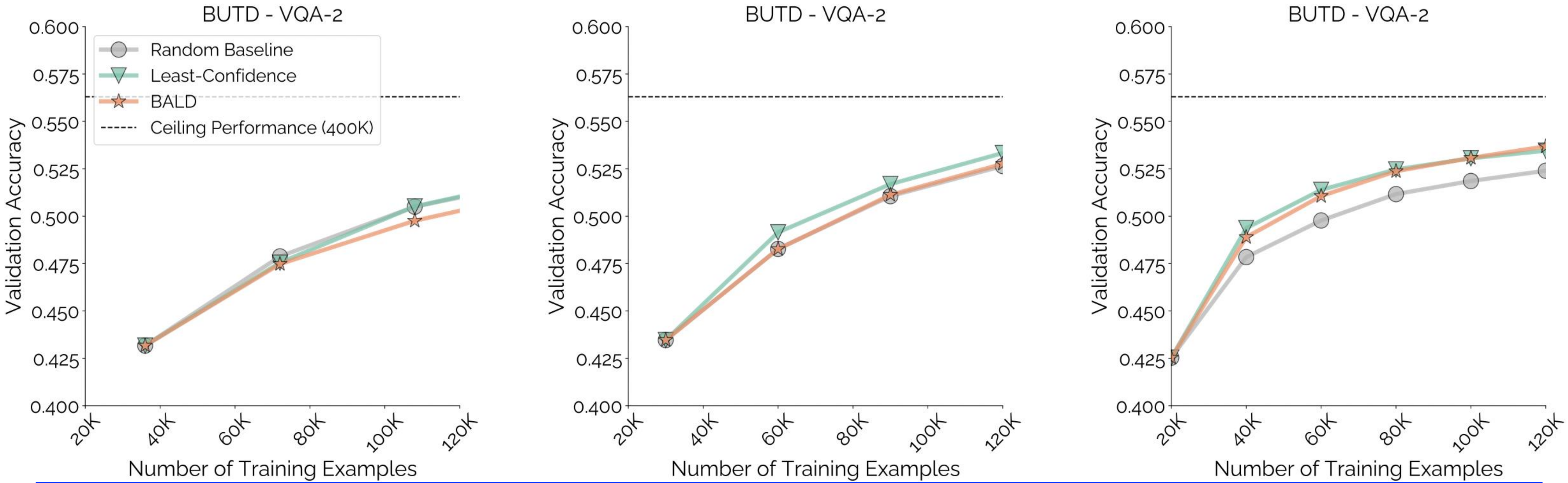


Figure 7: Example groups of collective outliers in the VQA-2 and GQA datasets.

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; e.g., (ICCV 2023 Papers)

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 Computer Vision

- Motivation
- Model Compression
- Curriculum Learning
- Active Learning
- Faculty Course Questionnaire (FCQ)


The End