Model Compression

Danna Gurari
University of Colorado Boulder
Fall 2023

https://home.cs.colorado.edu/~DrG/Courses/RecentAdvancesInComputerVision/AboutCourse.html

Review

- Last lecture on style transfer:
  - Problem
  - Applications
  - Neural Style Transfer Model
  - Evaluation Metrics
  - Autoencoder-Based Models
  - Other Approaches

- Assignments (Canvas):
  - Project outline due earlier today
  - Project presentation (poster and video) due in two weeks

- Questions?

Today’s Topics

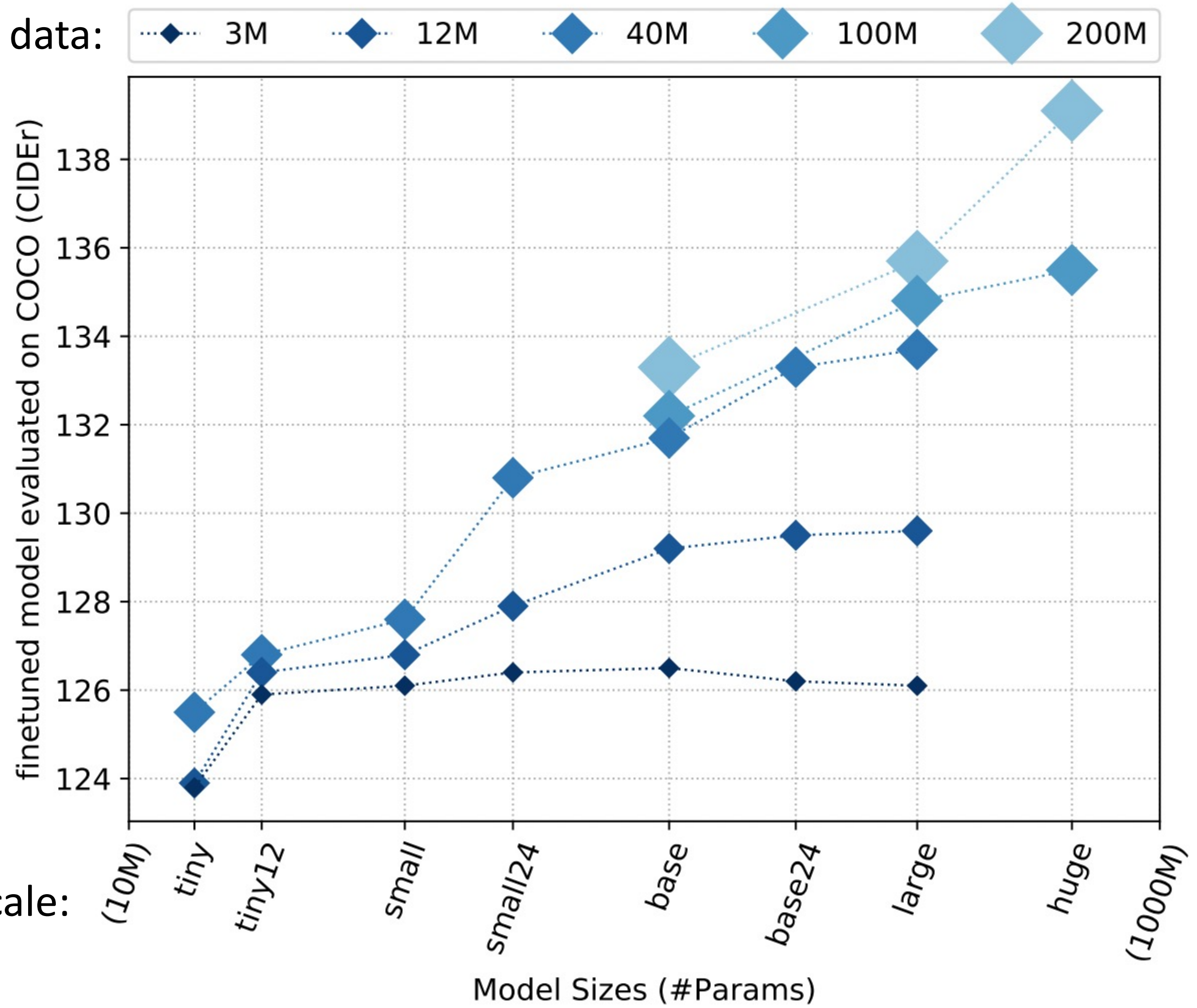
- Motivation
- Key idea: knowledge distillation (KD)
- Pioneering KD model for image classification
- Pioneering KD model for object detection
- State-of-the-art for KD (ICCV 2023 highlights)
- Interview about course: Ryan Layer

Today’s Topics

- 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)
- Interview about course: Ryan Layer

Trend: Parameter-Heavy Models; e.g.,

Amount of training data:



Logarithmic scale:

Larger models perform best (with lots of training data):

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

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

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

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

- Large inference time (i.e., incompatible for real-time applications)
- Large memory footprint (e.g., incompatible with limited memory on edge devices)
- Large computational cost (e.g., incompatible with limited battery on edge devices)
- Potential for large environmental costs

Idea: develop compact models so deep learning models can be used more efficiently and for more applications

Today’s Topics

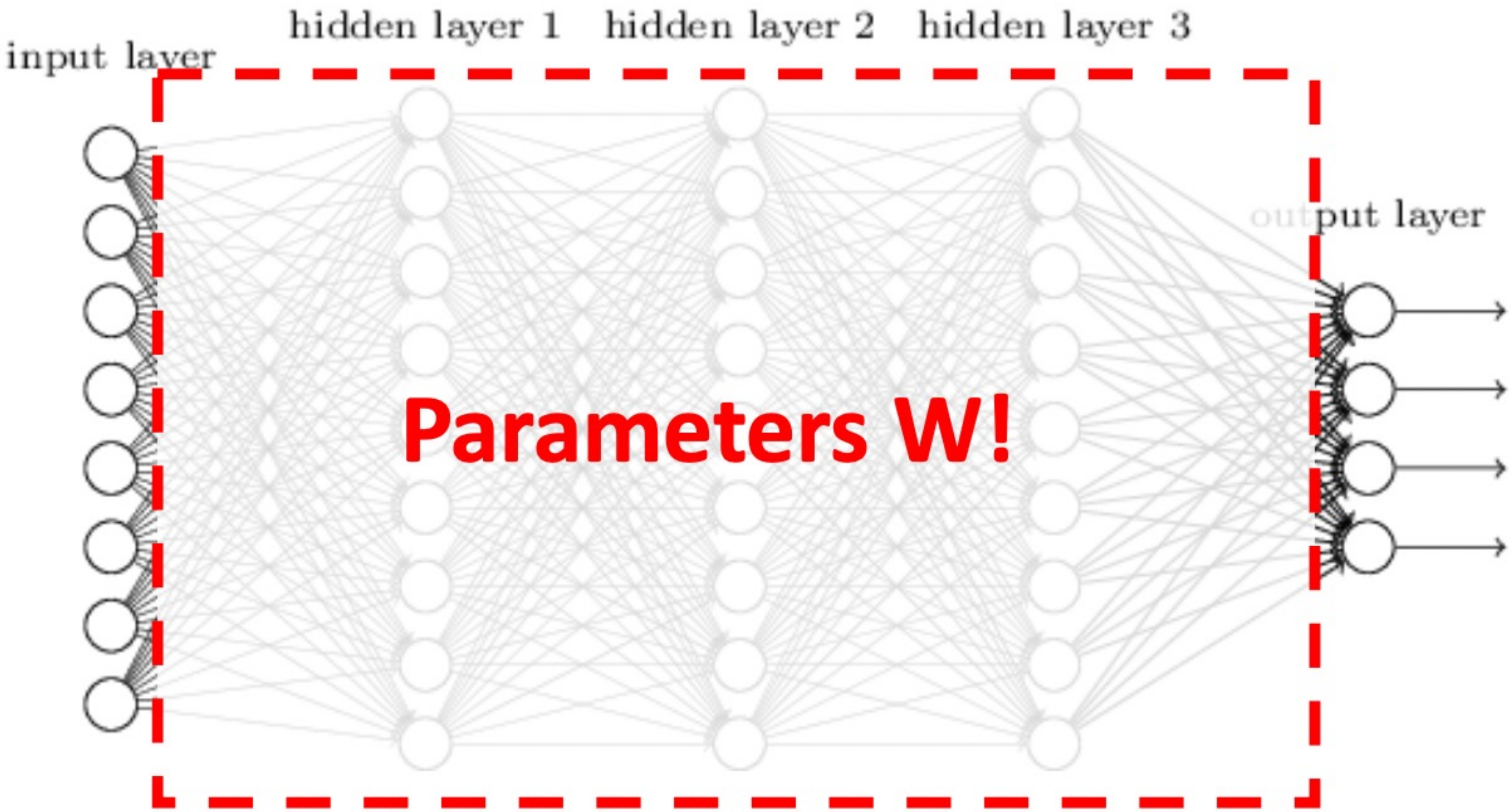
- Motivation
- Key idea: knowledge distillation (KD)
- Pioneering KD model for image classification
- Pioneering KD model for object detection
- State-of-the-art for KD (ICCV 2023 highlights)
- Interview about course: Ryan Layer

Popular Approach: Knowledge Distillation

A student learns from a knowledgeable teacher

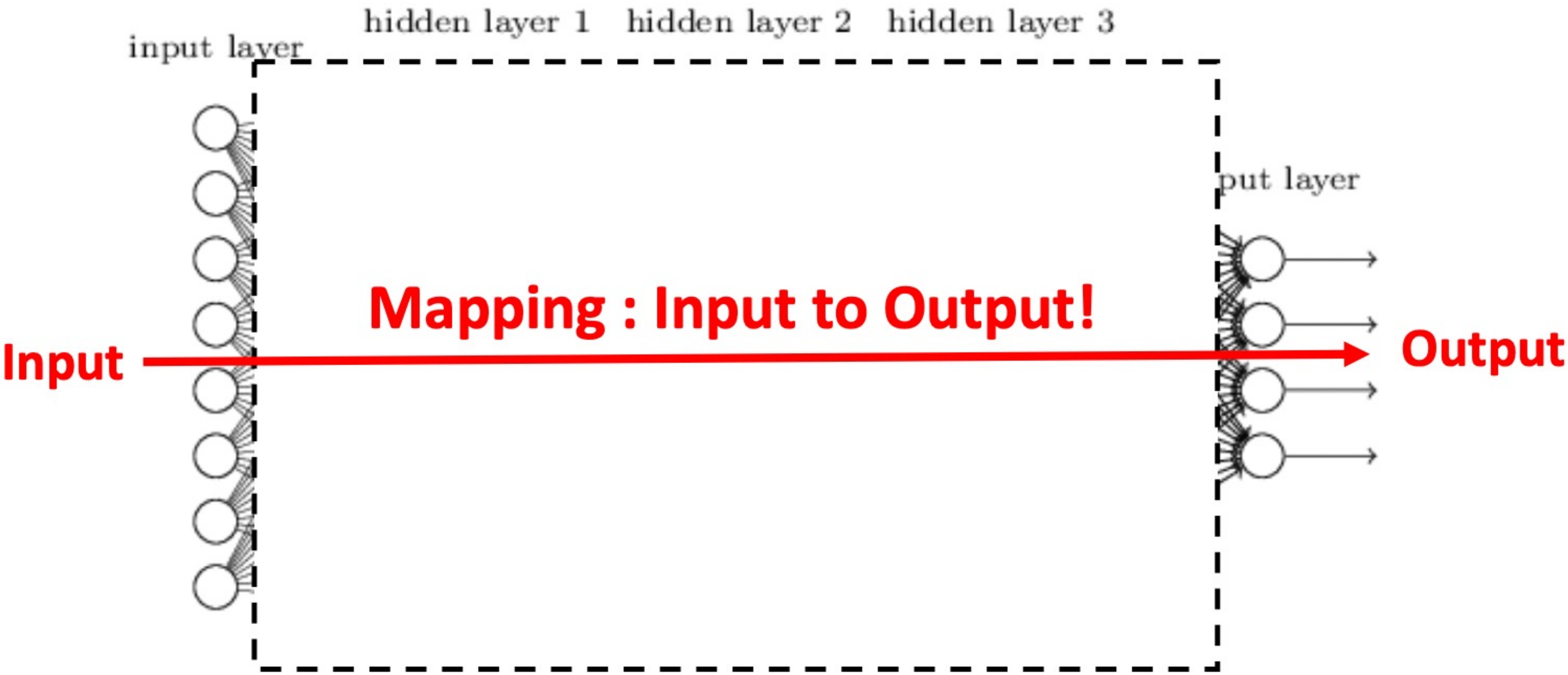
Image source: https://www.waterford.org/education/teacher-student-relationships/

Key Question: What is Knowledge?



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

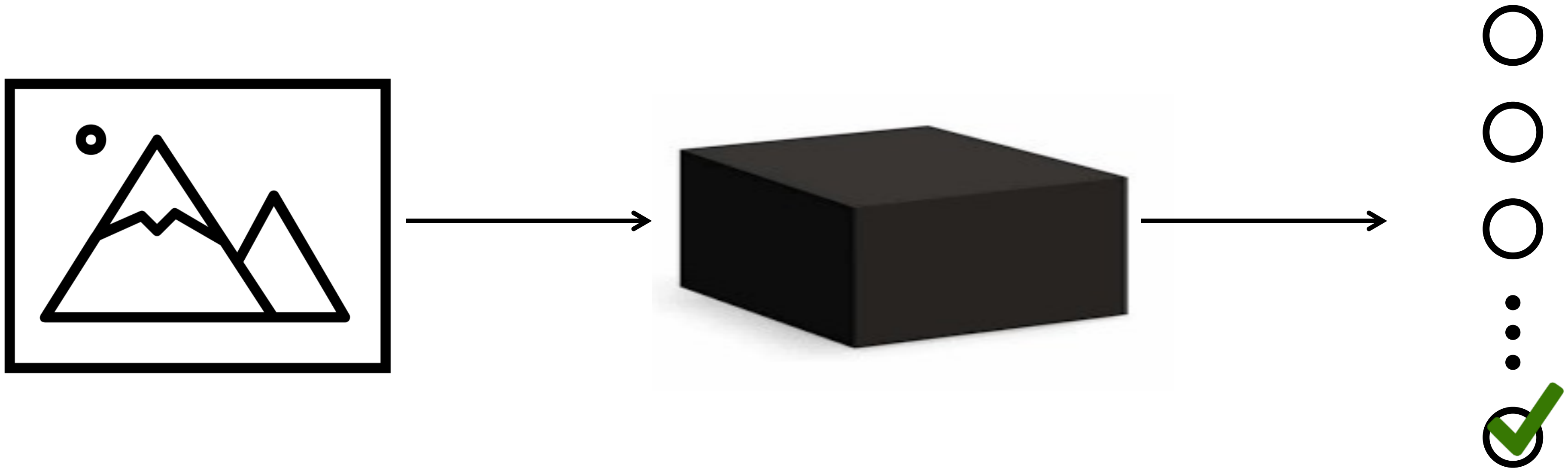
Knowledge Is: Input to Output Mapping



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

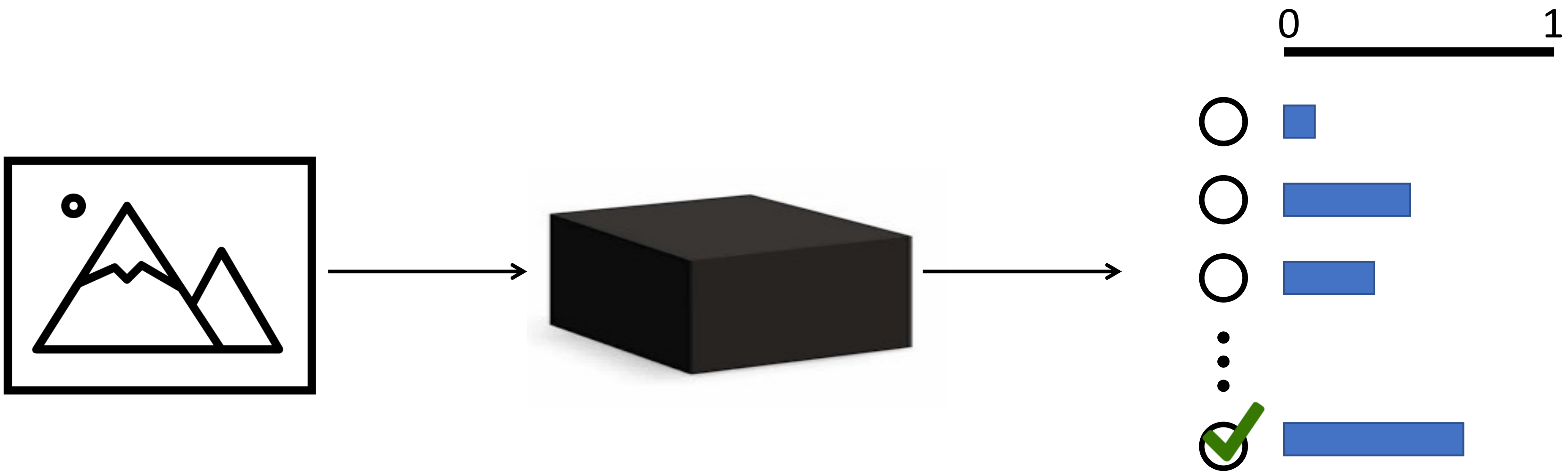
Knowledge Is: Input to Output Mapping

Target mapping: ground truth (1-hot vector)



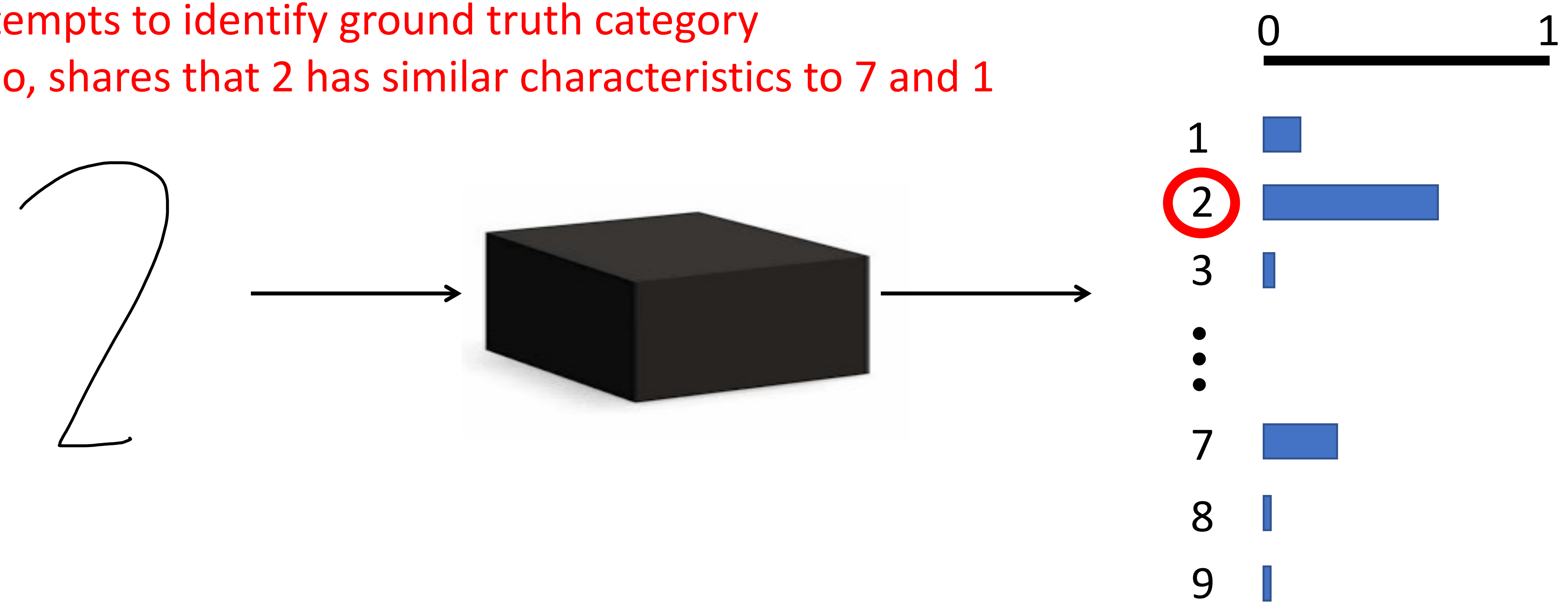
Knowledge Is: Input to Output Mapping

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



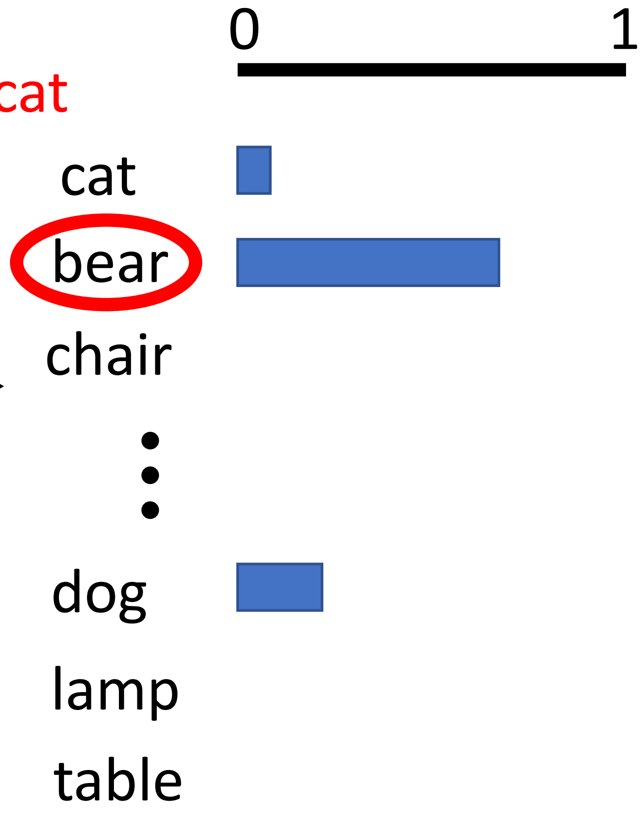
Knowledge Is: Input to Output Mapping

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



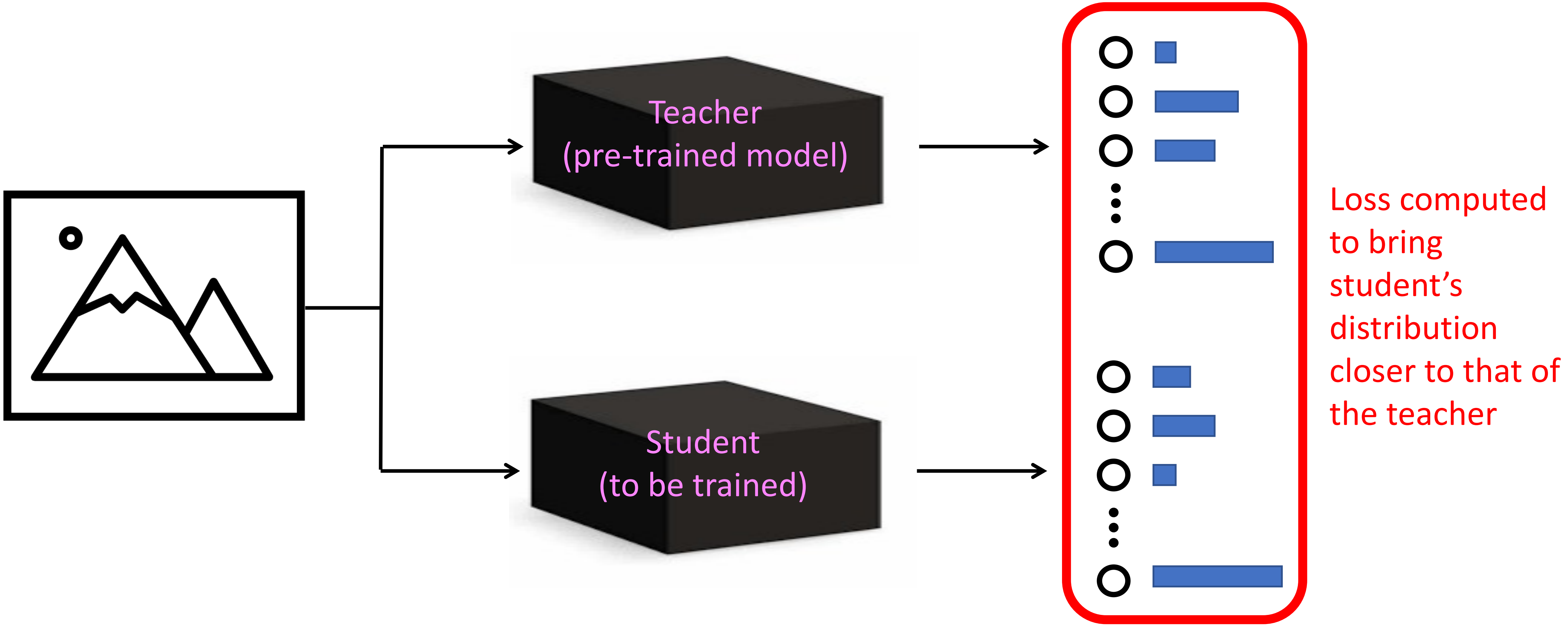
Knowledge Is: Input to Output Mapping

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

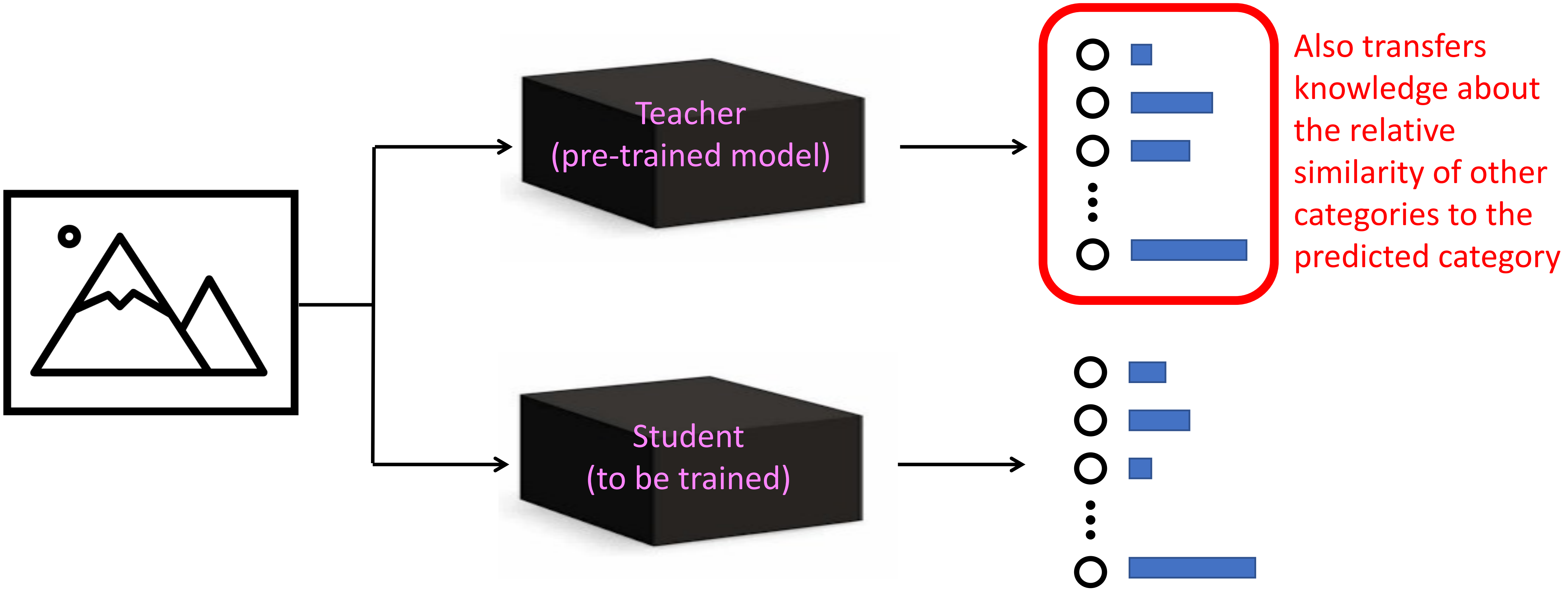
Idea: teach about ground truth and its relationships to other categories

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

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

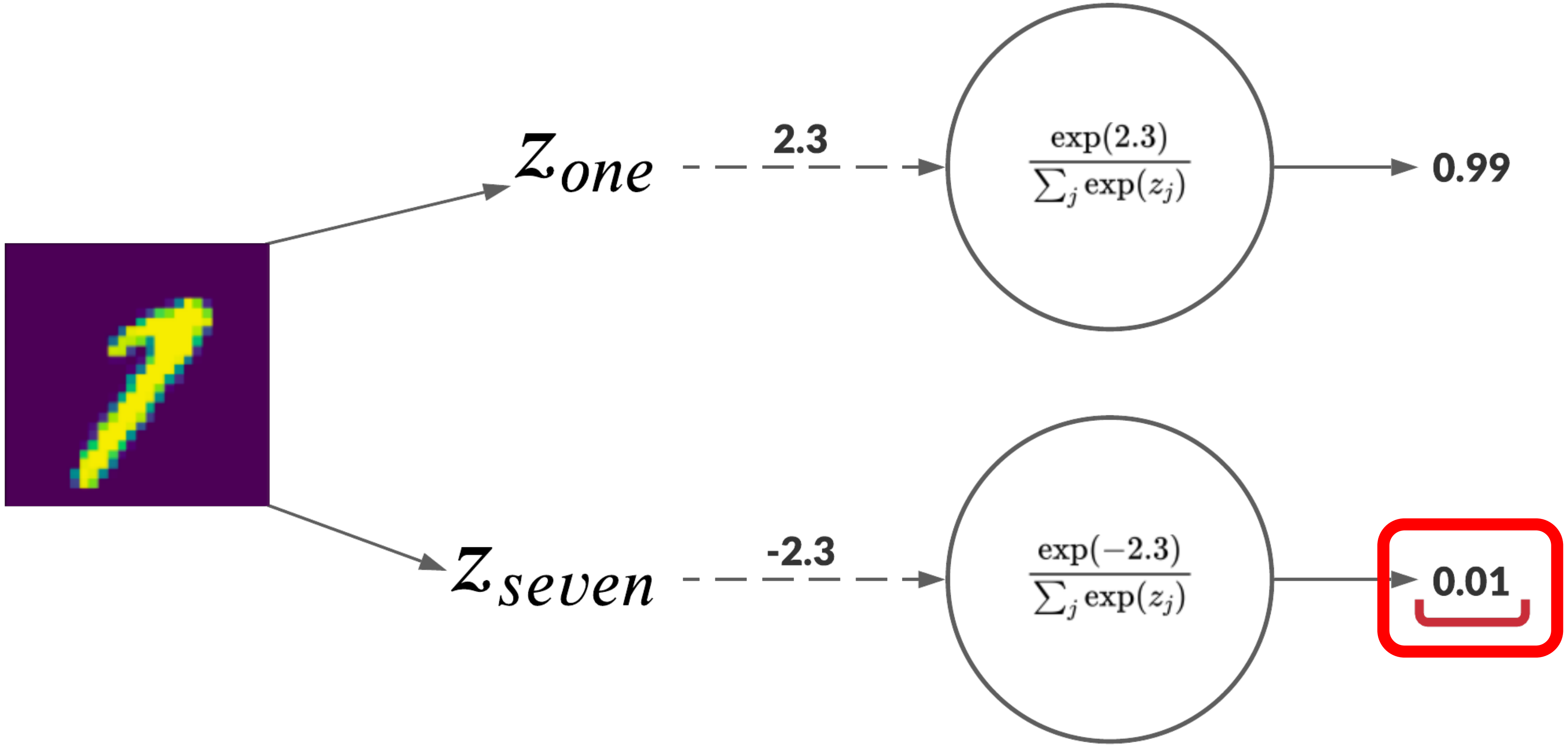


Knowledge Distillation: Teach Student the "Dark Knowledge" of Teacher



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

Recall Softmax: converts vector of 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

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

Get rid of negative values while preserving original order of scores

$$\sigma(\mathbf{z})_i = \frac{\boxed{e}^{\boxed{z_i}}}{\boxed{\sum_{j=1}^{K} \boxed{e}^{z_j}}}$$

i = 1, …, K

Number of classes

Divide each node's score by sum of all entries to make them sum to 1 (normalization)

Useful tutorial: https://towardsdatascience.com/exploring-the-softmax-function-578c8b0fb15

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

Generalized Softmax: converts vector of scores into a probability distribution that sums 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 vector of scores into a probability distribution that sums to 1 with temperature

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

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

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

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

Generalized Softmax: converts vector of scores into a probability distribution that sums to 1 with temperature

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

What is the effect of larger T values?

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

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

Generalized Softmax: converts vector of scores into a probability distribution that sums to 1 with temperature; e.g.,

$$\begin{bmatrix} 0.997 \\ 0.000 \\ 0.002 \\ 0.001 \\ 0.000 \end{bmatrix} \begin{array}{l} \text{Homework} \\ \text{Cake} \\ \text{Book} \\ \text{Assignment} \\ \text{Car} \end{array} \qquad \begin{bmatrix} 0.935 \\ 0.0001 \\ 0.046 \\ 0.017 \\ 0.0001 \end{bmatrix} \begin{array}{l} \text{Homework} \\ \text{Cake} \\ \text{Book} \\ \text{Assignment} \\ \text{Car} \end{array} \qquad \begin{bmatrix} 0.637 \\ 0.021 \\ 0.191 \\ 0.128 \\ 0.021 \end{bmatrix} \begin{array}{l} \text{Homework} \\ \text{Cake} \\ \text{Book} \\ \text{Assignment} \\ \text{Car} \end{array}$$

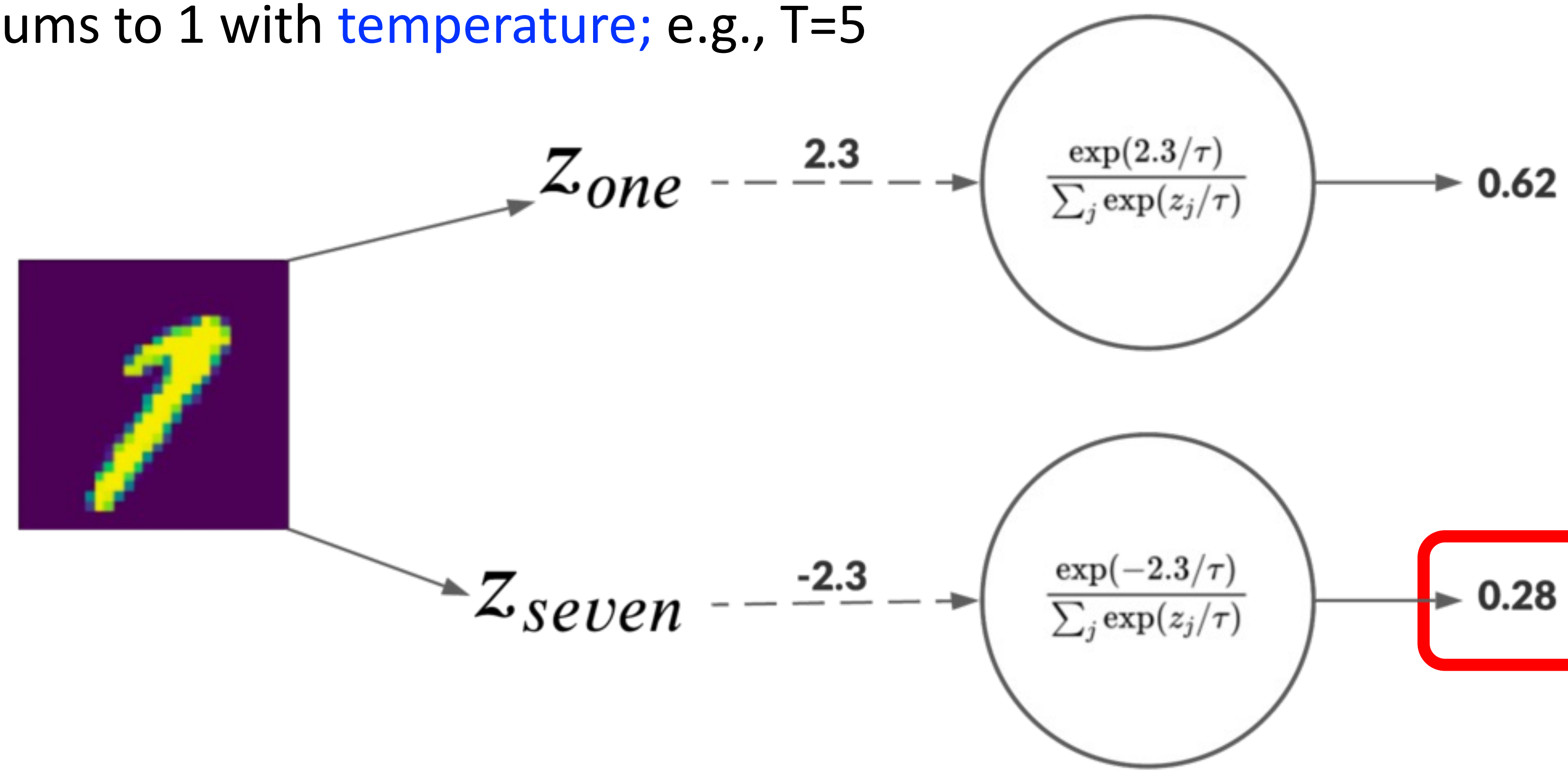
T=1 T=2 T=5

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

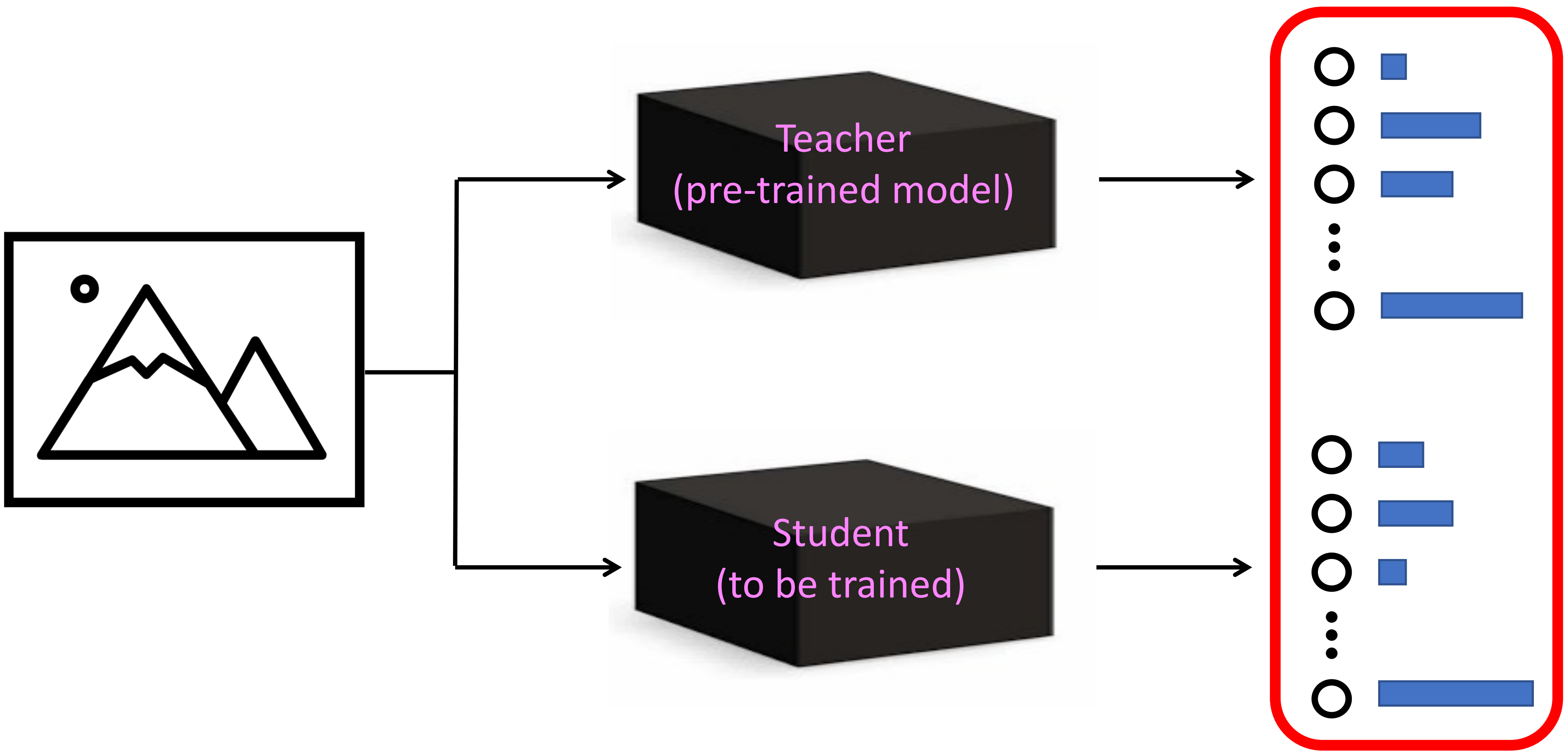
Generalized Softmax: converts vector of scores into a probability distribution that sums to 1 with temperature; e.g., T=5



Knowledge of this relationship enhanced from 0.01

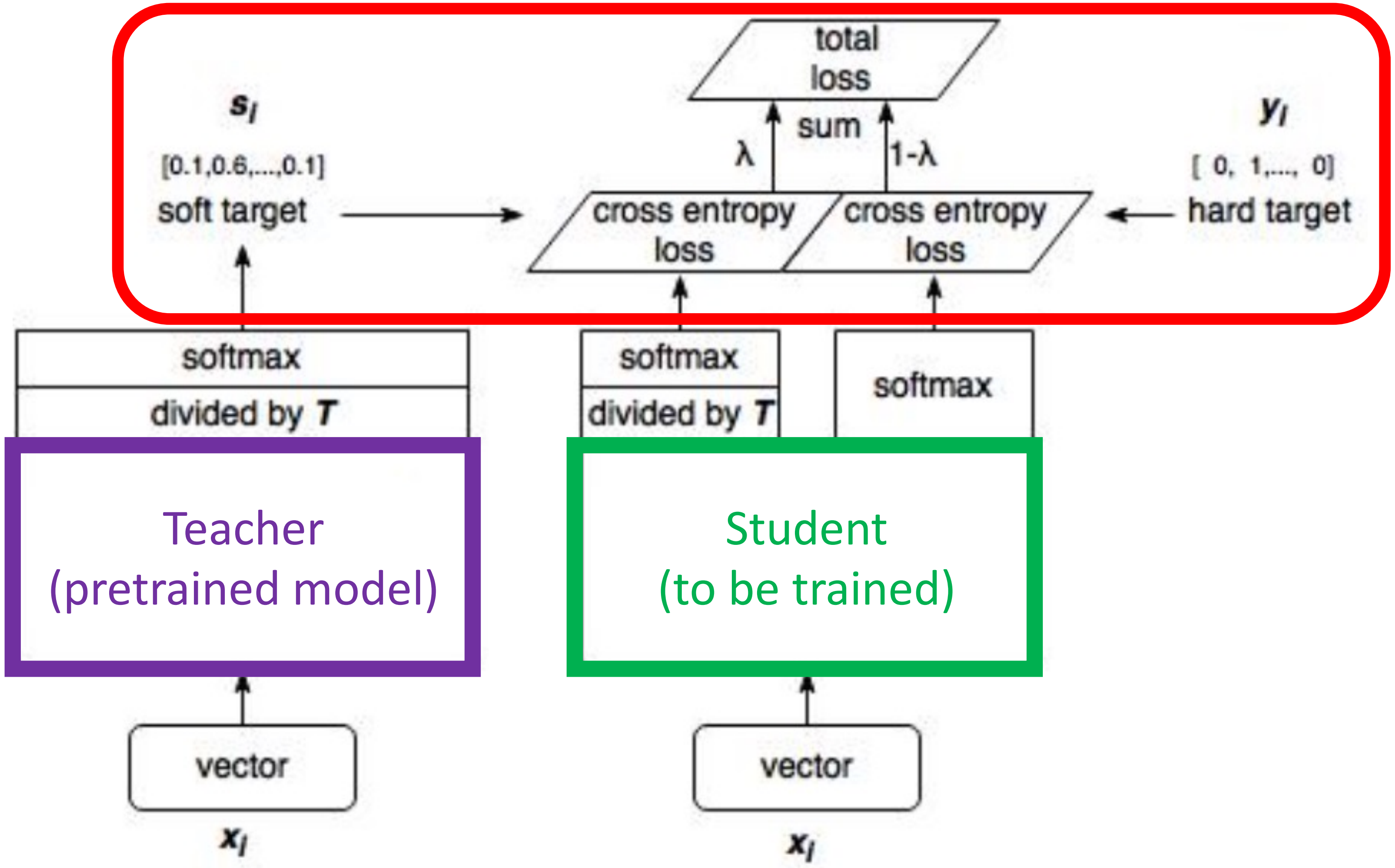
https://wandb.ai/authors/knowledge-distillation/reports/Distilling-Knowledge-in-Neural-Networks--VmlldzoyMjkxODk

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

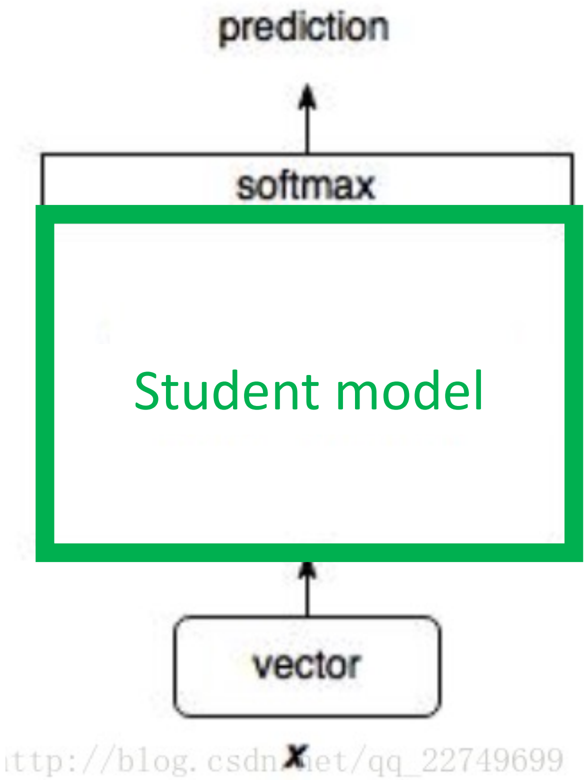
Knowledge Distillation: Teach Student the "Dark Knowledge" of Teacher



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

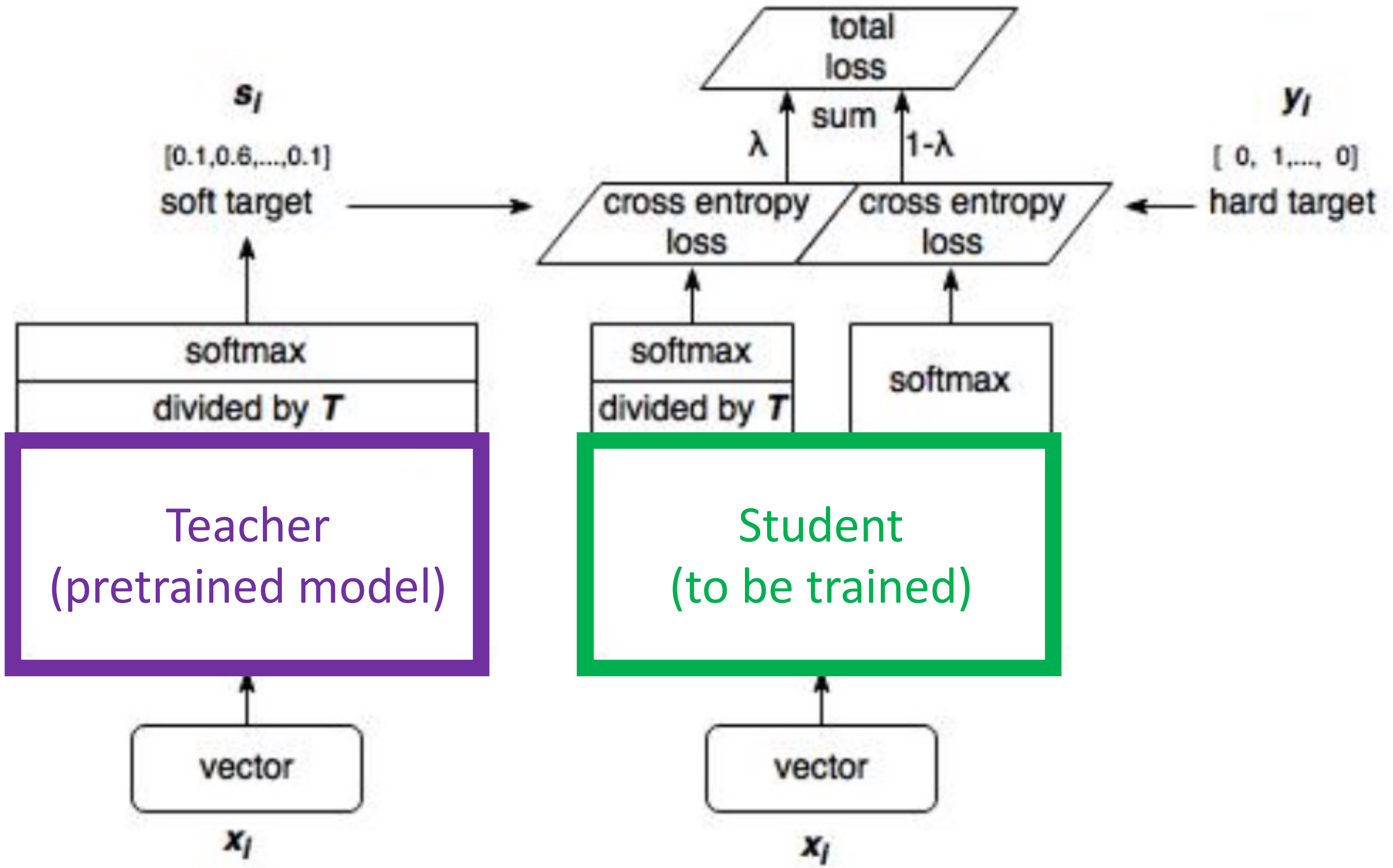
https://blog.csdn.net/qq_22749699/article/details/79460817

Knowledge Distillation: At Test Time



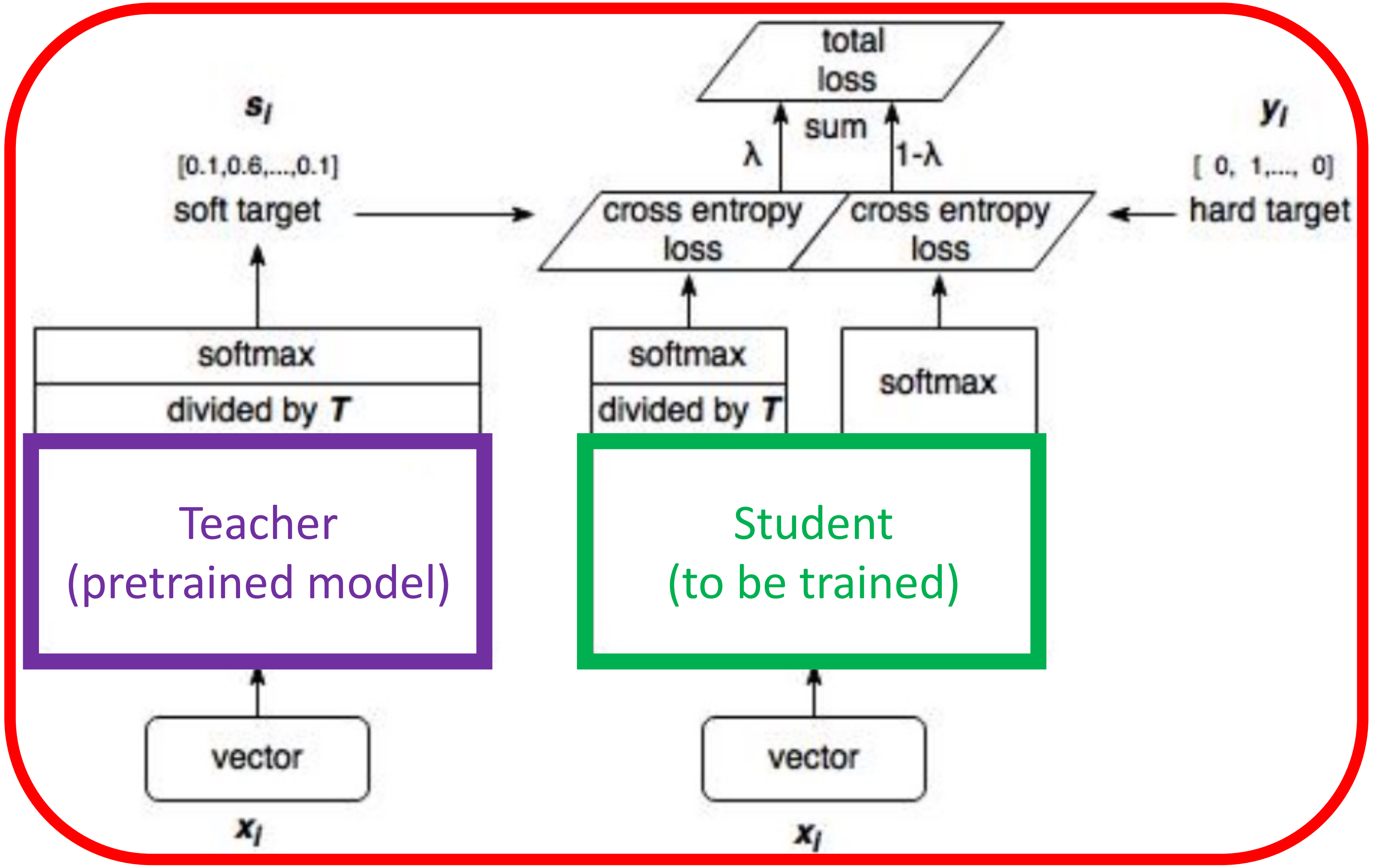
https://blog.csdn.net/qq_22749699/article/details/79460817

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

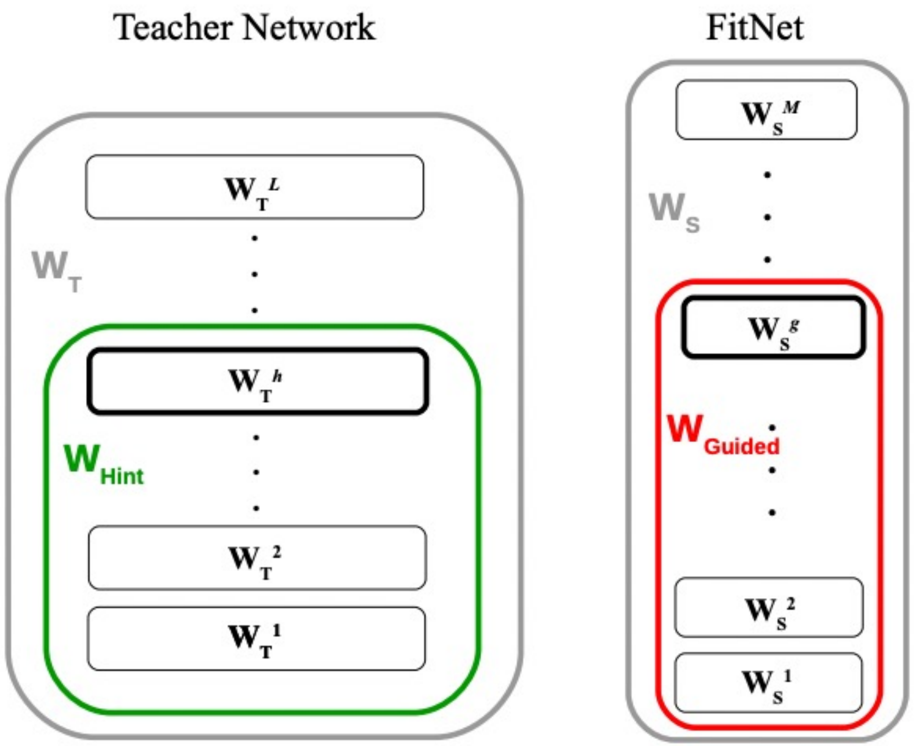


Knowledge distillation is a type of transfer learning

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



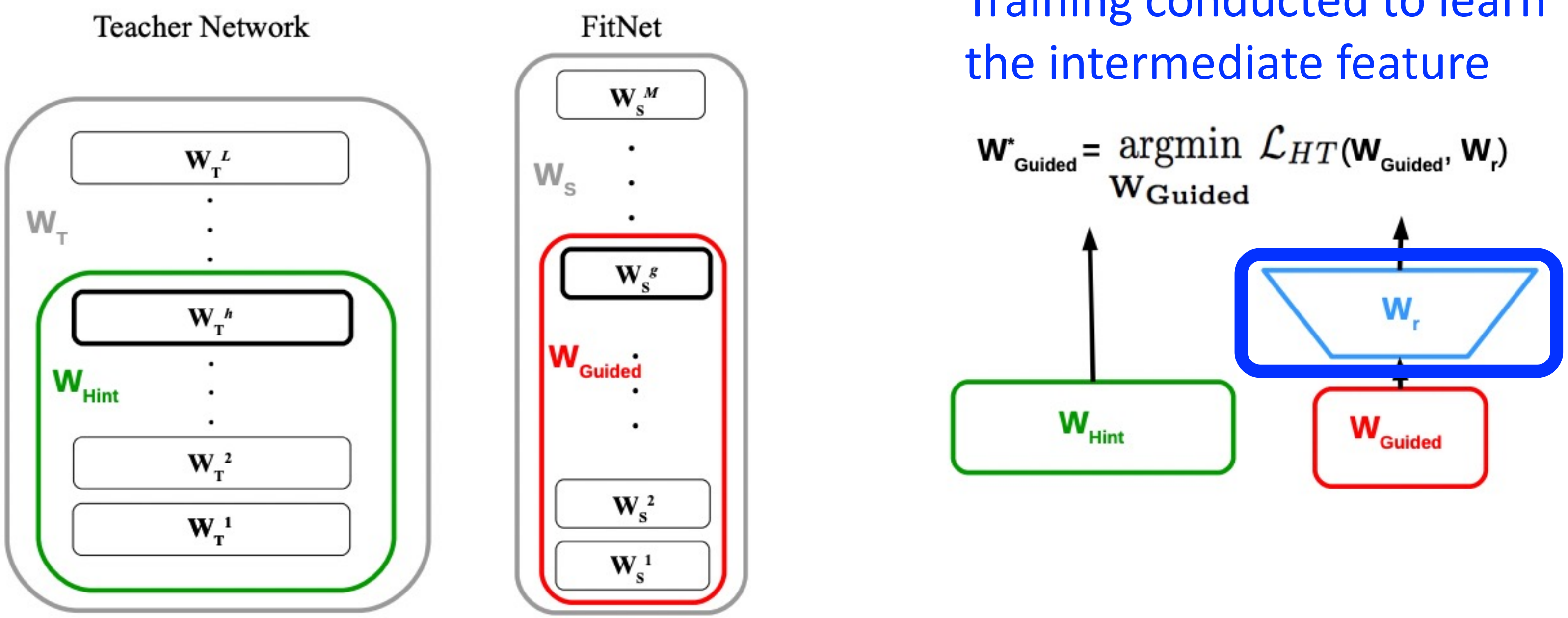
Romero et al. Fitnets: Hints for thin deep nets. ICLR 2015.

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

Training conducted to learn the intermediate feature

Layer added to match size of the hint's output layer



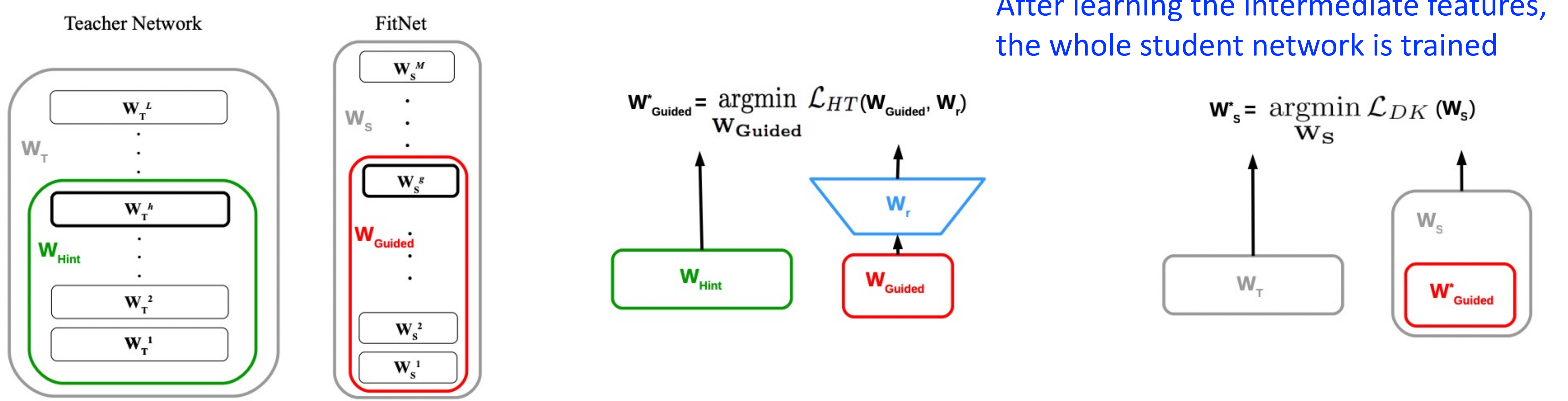
(a) Teacher and Student Networks (b) Hints Training

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.

Today’s Topics

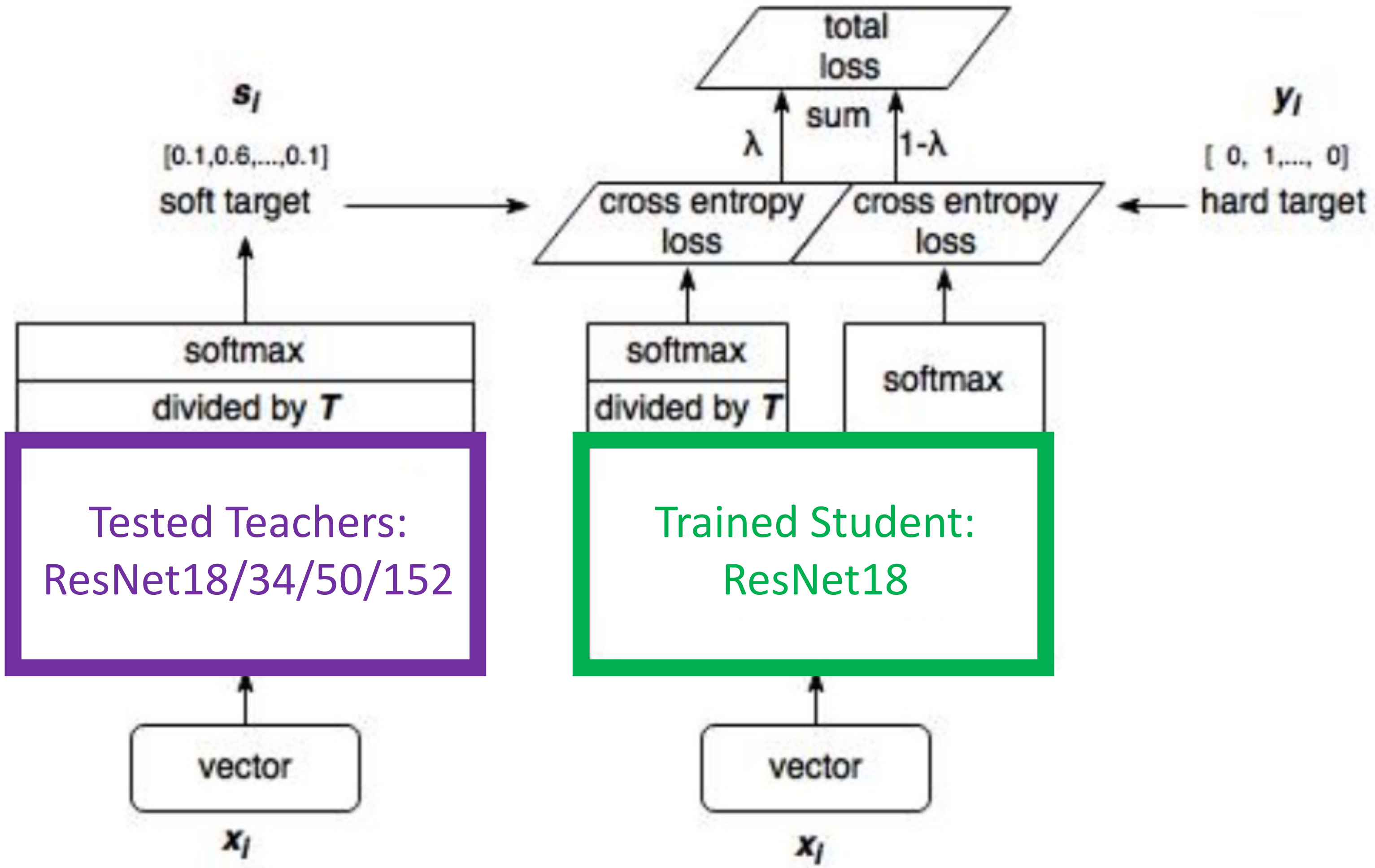
- Motivation
- Key idea: knowledge distillation (KD)
- Pioneering KD model for image classification
- Pioneering KD model for object detection
- State-of-the-art for KD (ICCV 2023 highlights)
- Interview about course: Ryan Layer

Recall Task: 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

Experiment: 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

Experiment: 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

Experiment: 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

Experiment: 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

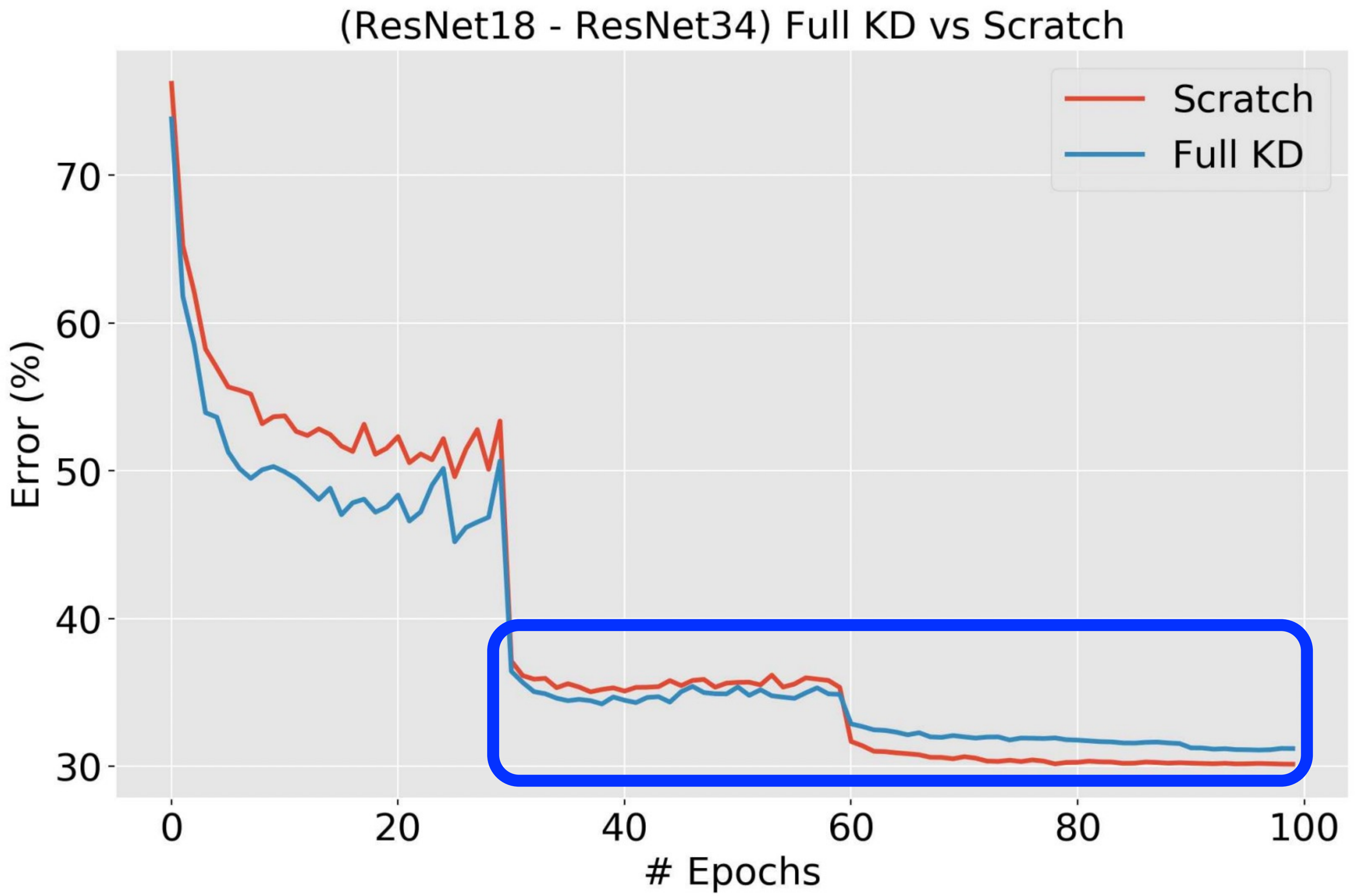
3. Student fails to accurately mimic teacher

Experimental analysis suggests this is the reason

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

Experiment: Why Might Students Fail to Mimic Teachers?

Hypothesis: student is underfitting because of lower capacity and so “minimizing one loss (KD loss) at the expense of the other (cross entropy loss)”

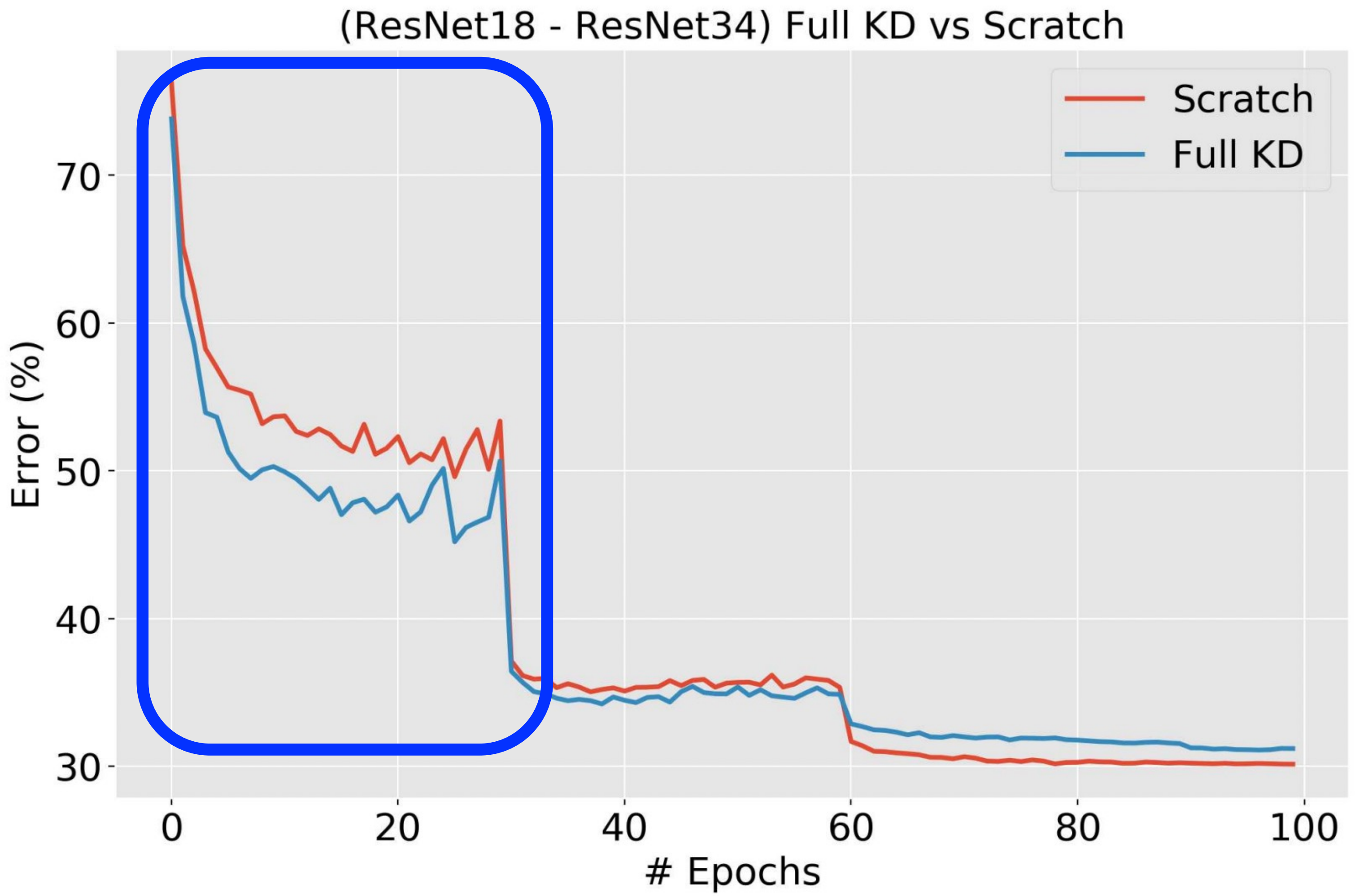


Cho and Hariharan. On the Efficacy of Knowledge Distillation. ICCV 2019

Experiment: 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

Experiments: 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

Experiments: Are Results from EKSD 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

Experiments: 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:

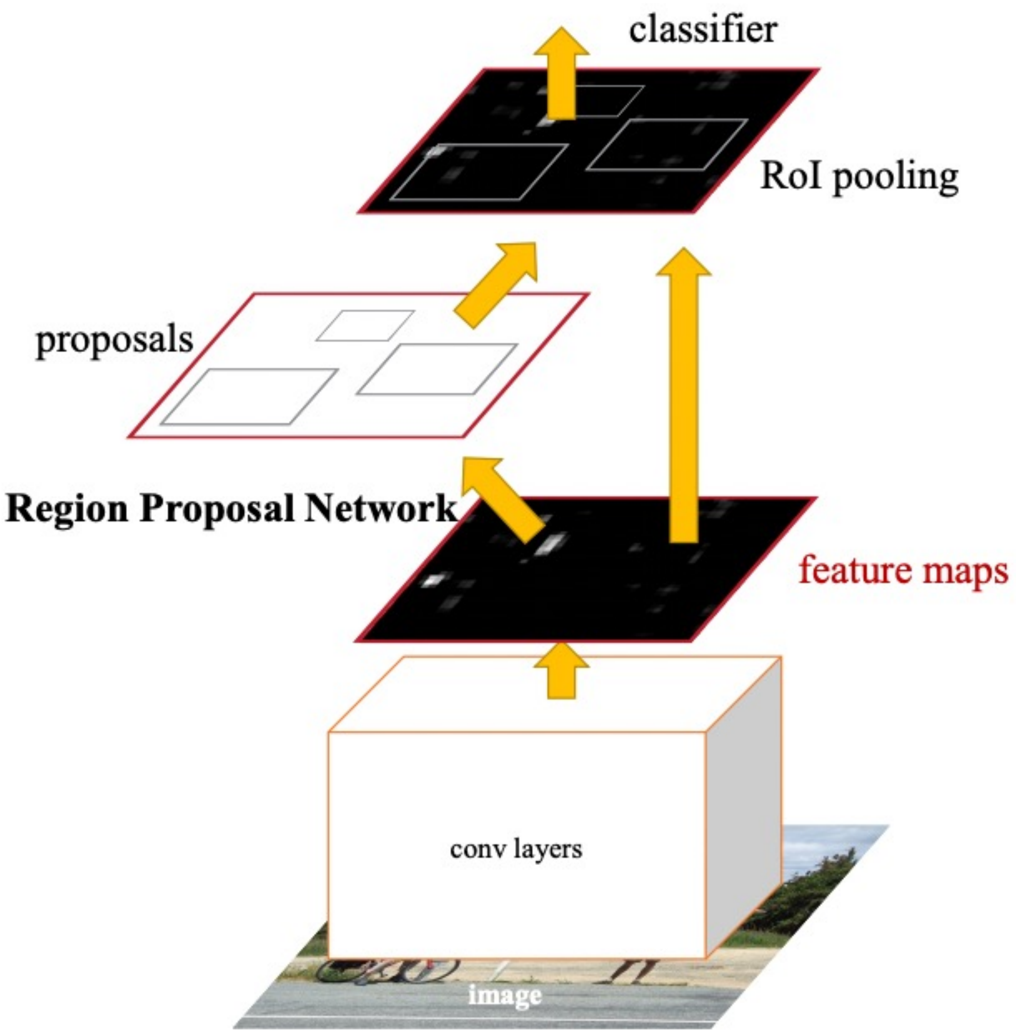
The student must be in the solution space of the teacher

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

Today’s Topics

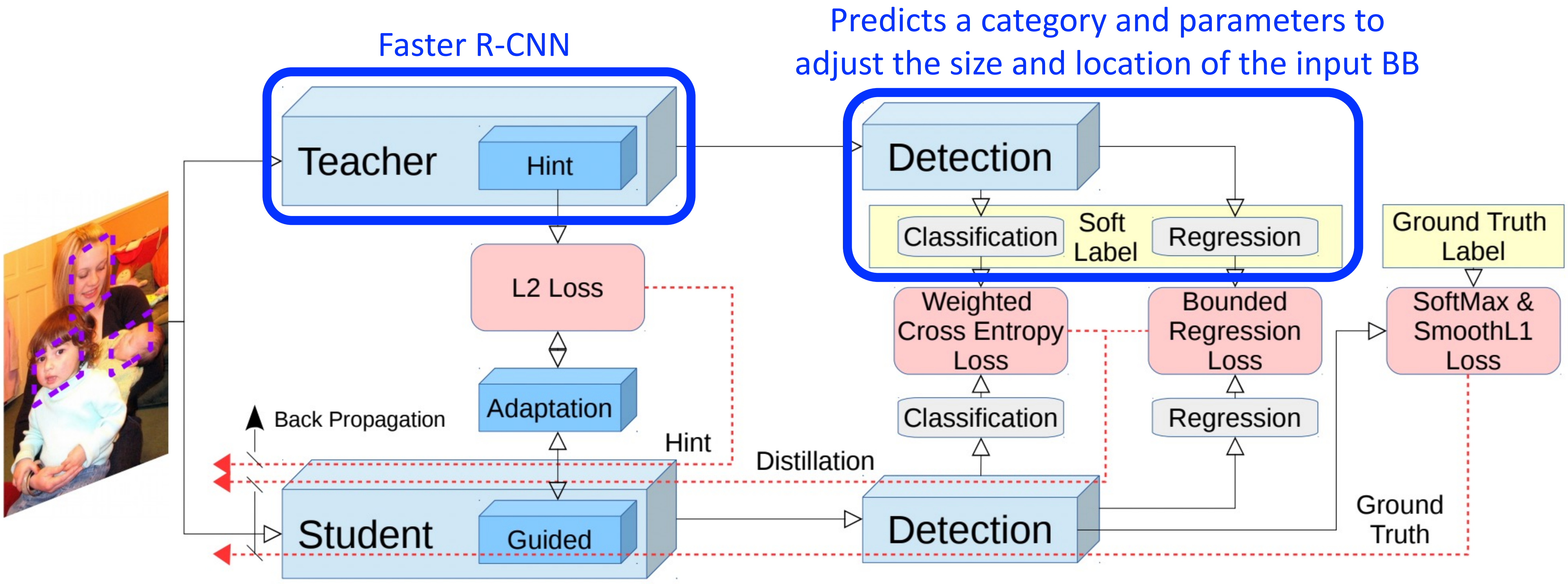
- Motivation
- Key idea: knowledge distillation (KD)
- Pioneering KD model for image classification
- Pioneering KD model for object detection
- State-of-the-art for KD (ICCV 2023 highlights)
- Interview about course: Ryan Layer

Recall Popular Detection Model: Faster R-CNN



Ren Shaoqing Ren et al. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." Neurips 2015.

Approach for Creating Compact Student Model



= backpropagation pathways

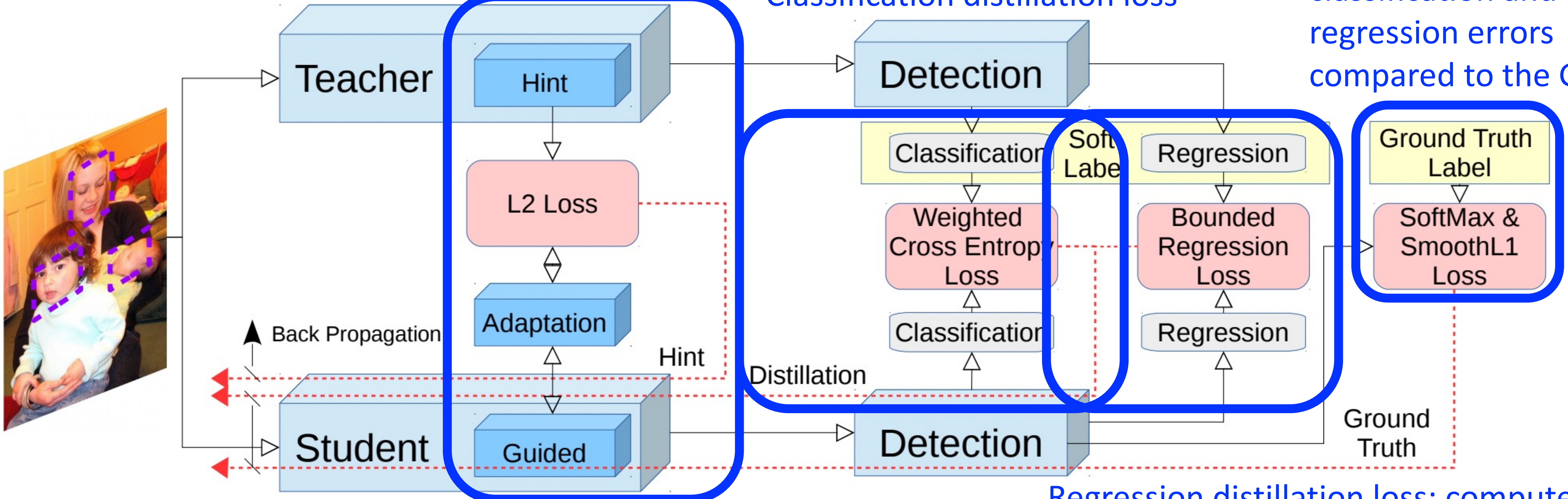
Chen et al. Learning efficient object detection models with knowledge distillation. Neurips 2017.

Approach for Creating Compact Student Model

A loss is computed to encourage the student's intermediate features to match those of the teacher

Classification distillation loss

Conventional loss computed for classification and regression errors compared to the GT



Regression distillation loss: computed if the student's distance to the GT exceeds the teacher's distance

= backpropagation pathways

Chen et al. Learning efficient object detection models with knowledge distillation. Neurips 2017.

Experiments

4 student models

3 teacher models

mAP scores for 5 datasets

Student	Model Info	Teacher	PASCAL	COCO@.5	COCO@[.5,.95]	KITTI	ILSVRC
Tucker	11M / 47ms	-	54.7	25.4	11.8	49.3	20.6
		AlexNet	57.6 (+2.9)	26.5 (+1.2)	12.3 (+0.5)	51.4 (+2.1)	23.6 (+1.3)
		VGGM	58.2 (+3.5)	26.4 (+1.1)	12.2 (+0.4)	51.4 (+2.1)	23.9 (+1.6)
		VGG16	59.4 (+4.7)	28.3 (+2.9)	12.6 (+0.8)	53.7 (+4.4)	24.4 (+2.1)
AlexNet	62M / 74ms	-	57.2	32.5	15.8	55.1	27.3
		VGGM	59.2 (+2.0)	33.4 (+0.9)	16.0 (+0.2)	56.3 (+1.2)	28.7 (+1.4)
		VGG16	60.1 (+2.9)	35.8 (+3.3)	16.9 (+1.1)	58.3 (+3.2)	30.1 (+2.8)
VGGM	80M / 86ms	-	59.8	33.6	16.1	56.7	31.1
		VGG16	63.7 (+3.9)	37.2 (+3.6)	17.3 (+1.2)	58.6 (+2.3)	34.0 (+2.9)
VGG16	138M / 283ms	-	70.4	45.1	24.2	59.2	35.6

params / speed

- means no distillation or, in other words, trained from scratch

What trends do you observe from these results?

Chen et al. Learning efficient object detection models with knowledge distillation. Neurips 2017.

Experiments

4 student models   3 teacher models   mAP scores for 5 datasets

Student	Model Info	Teacher	PASCAL	COCO@.5	COCO@[.5,.95]	KITTI	ILSVRC
Tucker	11M / 47ms	-	54.7	25.4	11.8	49.3	20.6
		AlexNet	57.6 (+2.9)	26.5 (+1.2)	12.3 (+0.5)	51.4 (+2.1)	23.6 (+1.3)
		VGGM	58.2 (+3.5)	26.4 (+1.1)	12.2 (+0.4)	51.4 (+2.1)	23.9 (+1.6)
		VGG16	59.4 (+4.7)	28.3 (+2.9)	12.6 (+0.8)	53.7 (+4.4)	24.4 (+2.1)
AlexNet	62M / 74ms	-	57.2	32.5	15.8	55.1	27.3
		VGGM	59.2 (+2.0)	33.4 (+0.9)	16.0 (+0.2)	56.3 (+1.2)	28.7 (+1.4)
		VGG16	60.1 (+2.9)	35.8 (+3.3)	16.9 (+1.1)	58.3 (+3.2)	30.1 (+2.8)
VGGM	80M / 86ms	-	59.8	33.6	16.1	56.7	31.1
		VGG16	63.7 (+3.9)	37.2 (+3.6)	17.3 (+1.2)	58.6 (+2.3)	34.0 (+2.9)
VGG16	138M / 283ms	-	70.4	45.1	24.2	59.2	35.6

- means no distillation or, in other words, trained from scratch

For all student-teacher pairs, knowledge distillation yields more compact, faster, and more accurate detections

Chen et al. Learning efficient object detection models with knowledge distillation. Neurips 2017.

Experiments

4 student models

3 teacher models

mAP scores for 5 datasets

Student	Model Info	Teacher	PASCAL	COCO@.5	COCO@[.5,.95]	KITTI	ILSVRC
Tucker	11M / 47ms	-	54.7	25.4	11.8	49.3	20.6
		AlexNet	57.6 (+2.9)	26.5 (+1.2)	12.3 (+0.5)	51.4 (+2.1)	23.6 (+1.3)
		VGGM	58.2 (+3.5)	26.4 (+1.1)	12.2 (+0.4)	51.4 (+2.1)	23.9 (+1.6)
		VGG16	59.4 (+4.7)	28.3 (+2.9)	12.6 (+0.8)	53.7 (+4.4)	24.4 (+2.1)
AlexNet	62M / 74ms	-	57.2	32.5	15.8	55.1	27.3
		VGGM	59.2 (+2.0)	33.4 (+0.9)	16.0 (+0.2)	56.3 (+1.2)	28.7 (+1.4)
		VGG16	60.1 (+2.9)	35.8 (+3.3)	16.9 (+1.1)	58.3 (+3.2)	30.1 (+2.8)
VGGM	80M / 86ms	-	59.8	33.6	16.1	56.7	31.1
		VGG16	63.7 (+3.9)	37.2 (+3.6)	17.3 (+1.2)	58.6 (+2.3)	34.0 (+2.9)
VGG16	138M / 283ms	-	70.4	45.1	24.2	59.2	35.6

- means no distillation or, in other words, trained from scratch

Larger teachers lead to greater performance improvements for distilled models

Chen et al. Learning efficient object detection models with knowledge distillation. Neurips 2017.

Experiments

4 student models

3 teacher models

mAP scores for 5 datasets

Student	Model Info	Teacher	PASCAL	COCO@.5	COCO@[.5,.95]	KITTI	ILSVRC
Tucker	11M / 47ms	-	54.7	25.4	11.8	49.3	20.6
		AlexNet	57.6 (+2.9)	26.5 (+1.2)	12.3 (+0.5)	51.4 (+2.1)	23.6 (+1.3)
		VGGM	58.2 (+3.5)	26.4 (+1.1)	12.2 (+0.4)	51.4 (+2.1)	23.9 (+1.6)
		VGG16	59.4 (+4.7)	28.3 (+2.9)	12.6 (+0.8)	53.7 (+4.4)	24.4 (+2.1)
AlexNet	62M / 74ms	-	57.2	32.5	15.8	55.1	27.3
		VGGM	59.2 (+2.0)	33.4 (+0.9)	16.0 (+0.2)	56.3 (+1.2)	28.7 (+1.4)
		VGG16	60.1 (+2.9)	35.8 (+3.3)	16.9 (+1.1)	58.3 (+3.2)	30.1 (+2.8)
VGGM	80M / 86ms	-	59.8	33.6	16.1	56.7	31.1
		VGG16	63.7 (+3.9)	37.2 (+3.6)	17.3 (+1.2)	58.6 (+2.3)	34.0 (+2.9)
VGG16	138M / 283ms	-	70.4	45.1	24.2	59.2	35.6

\- means no distillation or, in other words, trained from scratch

Why do you think there are performance improvements from model compression?

Chen et al. Learning efficient object detection models with knowledge distillation. Neurips 2017.

Experiments

mAP scores for 5 datasets

Student	Model Info	Teacher	PASCAL	COCO@.5	COCO@[.5,.95]	KITTI	ILSVRC
Tucker	11M / 47ms	-	54.7	25.4	11.8	49.3	20.6
		AlexNet	57.6 (+2.9)	26.5 (+1.2)	12.3 (+0.5)	51.4 (+2.1)	23.6 (+1.3)
		VGGM	58.2 (+3.5)	26.4 (+1.1)	12.2 (+0.4)	51.4 (+2.1)	23.9 (+1.6)
		VGG16	59.4 (+4.7)	28.3 (+2.9)	12.6 (+0.8)	53.7 (+4.4)	24.4 (+2.1)
AlexNet	62M / 74ms	-	57.2	32.5	15.8	55.1	27.3
		VGGM	59.2 (+2.0)	33.4 (+0.9)	16.0 (+0.2)	56.3 (+1.2)	28.7 (+1.4)
		VGG16	60.1 (+2.9)	35.8 (+3.3)	16.9 (+1.1)	58.3 (+3.2)	30.1 (+2.8)
VGGM	80M / 86ms	-	59.8	33.6	16.1	56.7	31.1
		VGG16	63.7 (+3.9)	37.2 (+3.6)	17.3 (+1.2)	58.6 (+2.3)	34.0 (+2.9)
VGG16	138M / 283ms	-	70.4	45.1	24.2	59.2	35.6

- means no distillation or, in other words, trained from scratch

Still, larger models with more parameters return the best results.

Chen et al. Learning efficient object detection models with knowledge distillation. Neurips 2017.

Today’s Topics

- 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)
- Interview about course: Ryan Layer

ICCV 2023 – 19 Papers with KD in Title; e.g.,

Label-Guided Knowledge Distillation for Continual Semantic Segmentation

Wang[1],
eng Lin[1*]
al University, Singapore
(HMGICS)
com, gslin@ntu.edu.sg

UniKD: Universa

Shanshan
[1] Tsinghua

Remembering Normality: Memory-guided Knowledge Distillation for Unsupervised Anomaly Detection

Zhihao Gu[1*], Liang Liu[2*], Xu Chen[2*], Ran Yi[1], Jiangning Zhang[2],
Yabiao Wang[2], Chengjie Wang[1,2], Annan Shu[3], Guannan Jiang[3], Lizhuang Ma[1†]
[1]Shanghai Jiao Tong University, China [2]Tencent YouTu Lab, China [3]CATL, China

Beyond the limitati

[2]UC Santa Cruz

Dual Learning with Dynamic Knowledge Distillation for Partially Relevant Video Retrieval

Jianfeng Dong[1,2], Minsong Zhang[1*], Zheng Zhang[1*], Xiank
Daizong Liu[3], Xiaoye Qu[4], Xun Wang[1,2], Baolong Li
[1]Zhejiang Gongshang University, [2]Zhejiang Key Lab of E-C
[3]Peking University, [4]Huazhong University of Science and T
https://github.com/HuiGuanLab/DL-DKD

Class-relation Knowledge Distillation for Novel Class Discovery

Peiyan Gu[1,*] Chuyu Zhang[1,2,*] Ruijie Xu[1] Xuming He[1,3]
[1]ShanghaiTech University, Shanghai, China [2]Lingang Laboratory, Shanghai, China
[3]Shanghai Engineering Research Center of Intelligent Vision and Imaging, Shanghai, China
{zhangchy2,gupy,xurj2022,hexm}@shanghaitech.edu.cn

What’s New with Knowledge Distillation?

- Ways to support many types of intermediate features for many models
- Enables efficient knowledge transfer by training new models with decontaminated information (more on efficient learning next lecture)

Today’s Topics

- Motivation
- Key idea: knowledge distillation (KD)
- Pioneering KD model for image classification
- Pioneering KD model for object detection
- State-of-the-art for KD (CVPR 2023 highlights)
- Interview about course: Ryan Layer


The End