Model Compression

Danna Gurari
University of Colorado Boulder
Spring 2025

https://dannagurari.colorado.edu/course/neural-networks-and-deep-learning-spring-2025/

Review

- Last lecture: Practical Systems-Level Development Challenge
  - Motivation
  - Data curation
  - Model maintenance
  - TAs' experiences

- Assignments (Canvas):
  - Lab assignment 3 grades out
  - Project outline due in 1.5 weeks

- Questions?

Today’s Topics

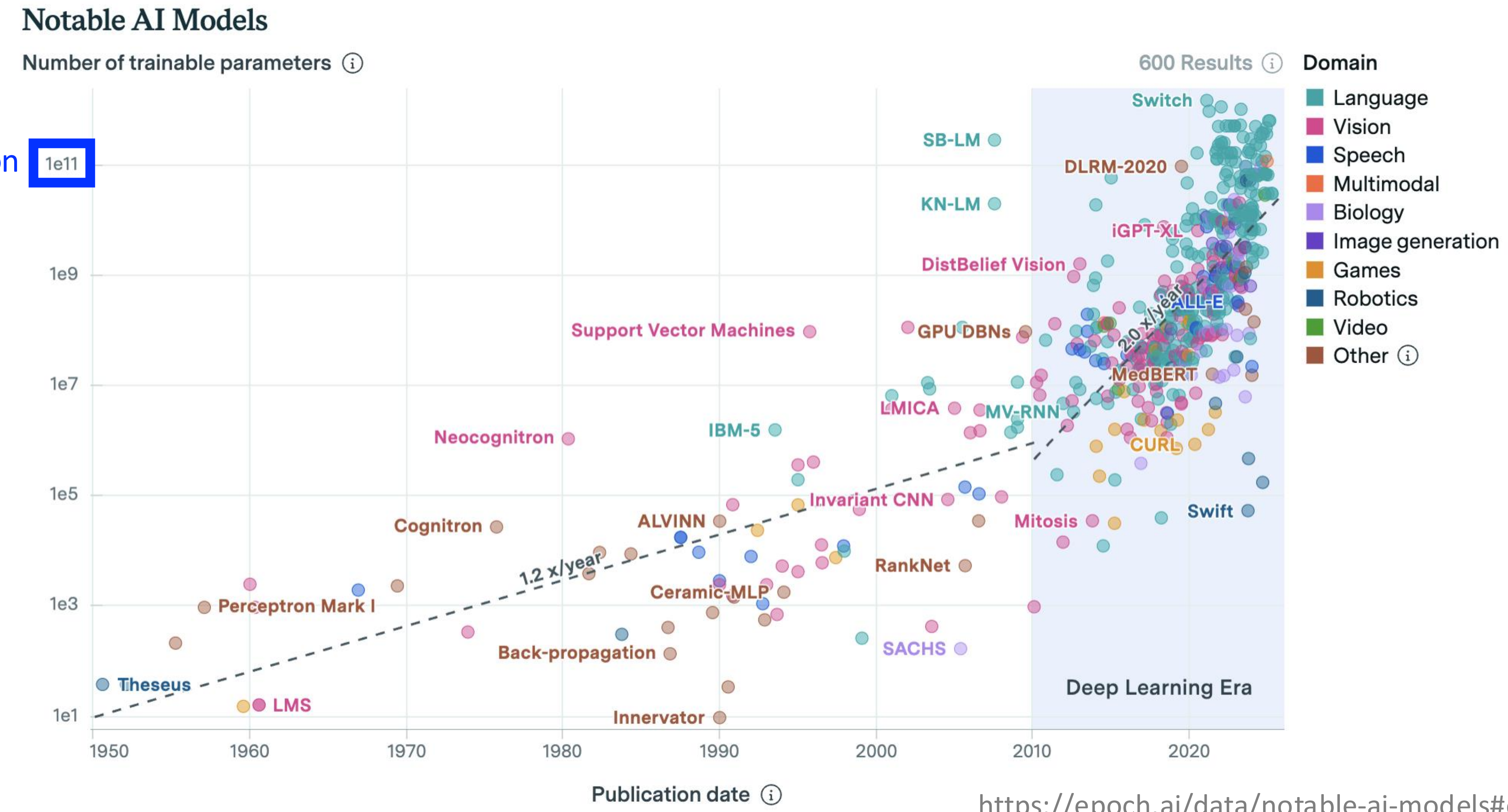
- Motivation
- Pruning
- Knowledge distillation
- Final project report: Overleaf tutorial

Today’s Topics

- Motivation
- Pruning
- Knowledge distillation
- Final project report: Overleaf tutorial

Trend: Parameter-Heavy Models (Scaling Law)

100 billion



https://epoch.ai/data/notable-ai-models#explore-the-data

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

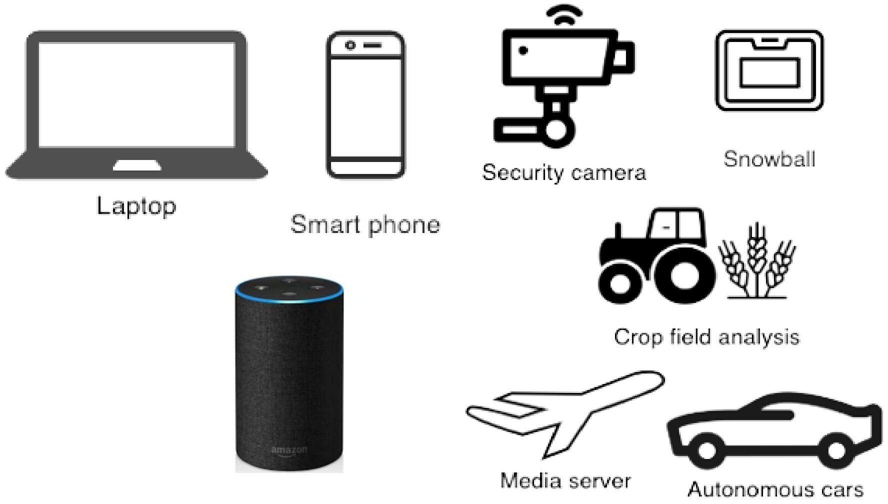
https://en.wikipedia.org/wiki/GPT-4

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



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

- Large inference time incompatible for real-time applications
- Large memory footprint incompatible for limited memory devices
- Large computational cost incompatible for limited battery devices
- Larger environmental costs

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

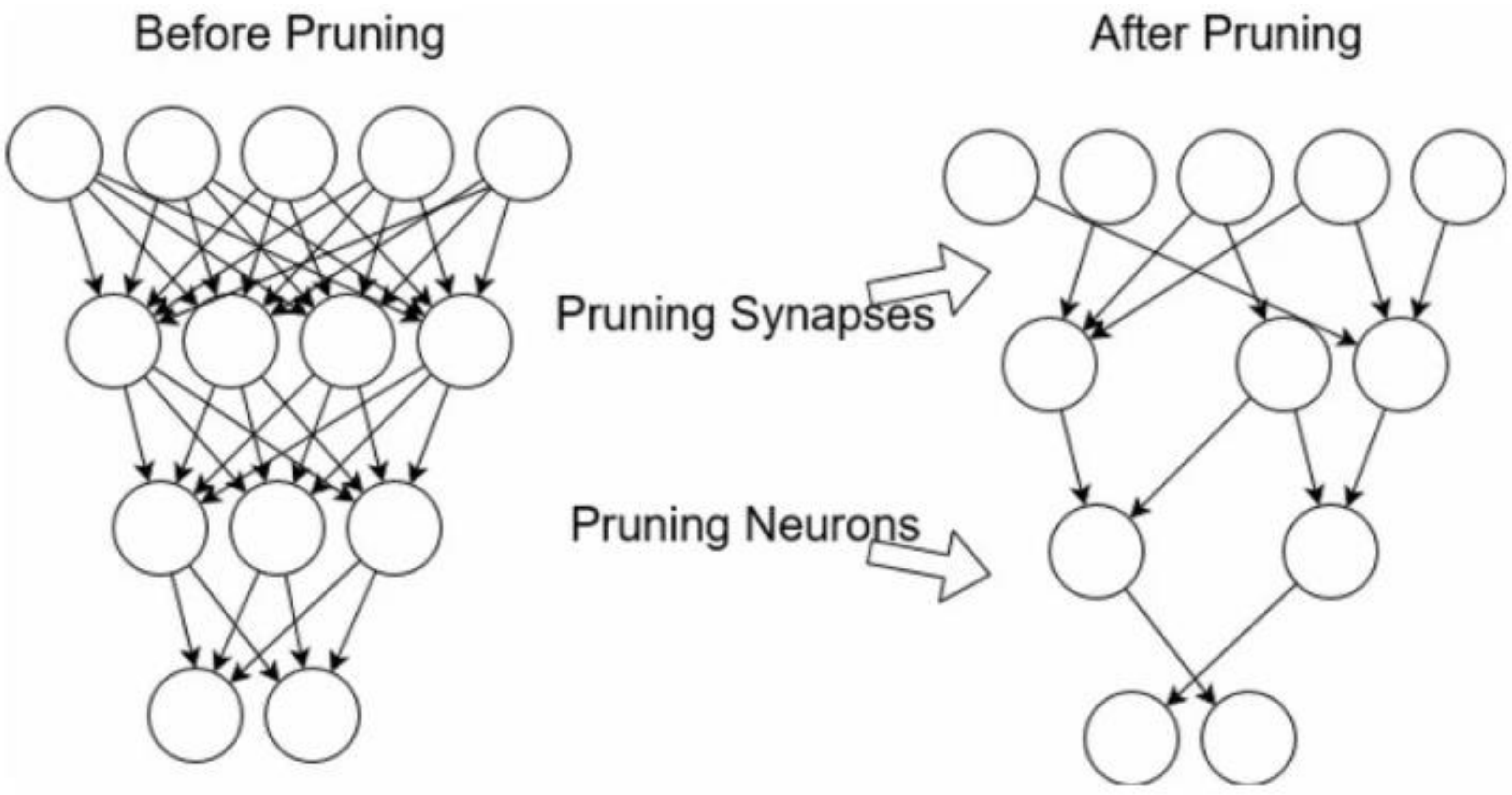
Today’s Topics

- Motivation
- Pruning
- Knowledge distillation
- Final project report: Overleaf tutorial

Idea: Trim Model to Necessary Content

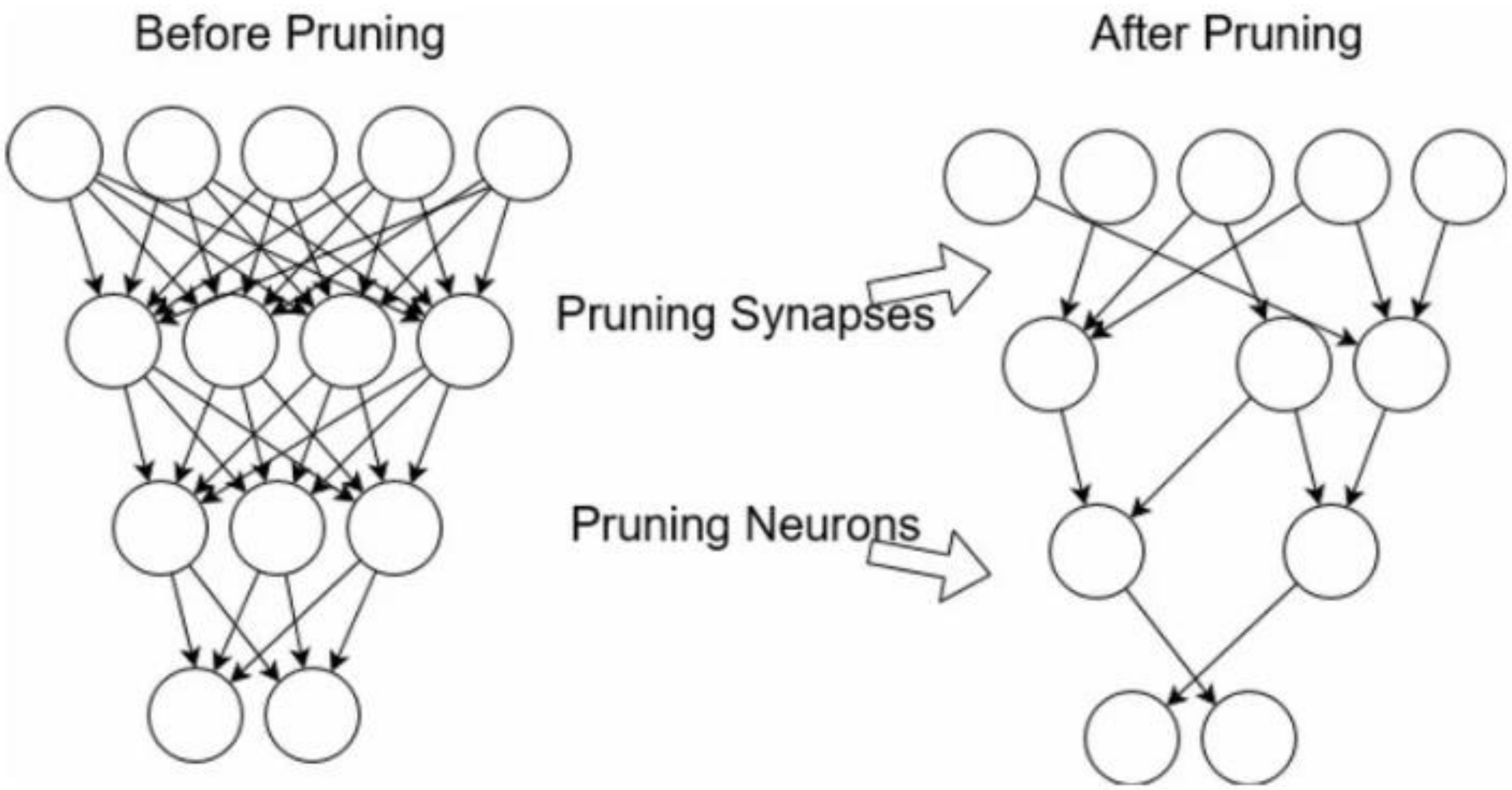
https://en.wikipedia.org/wiki/Topiary:

Pruning



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

Pruning: When?

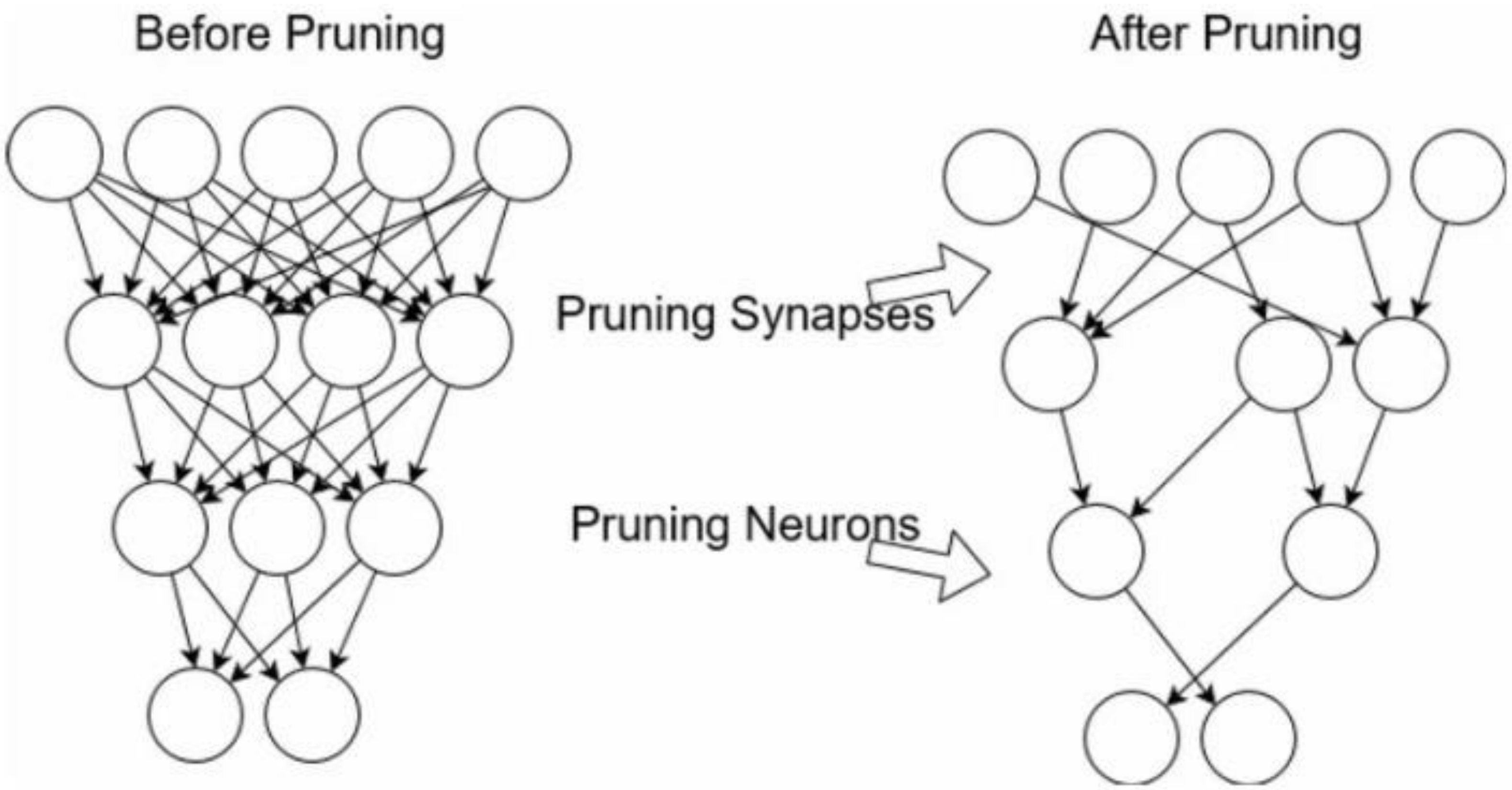


Can remove parameters:
- during training
- after training

Only former requires training a model

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

Pruning: What?

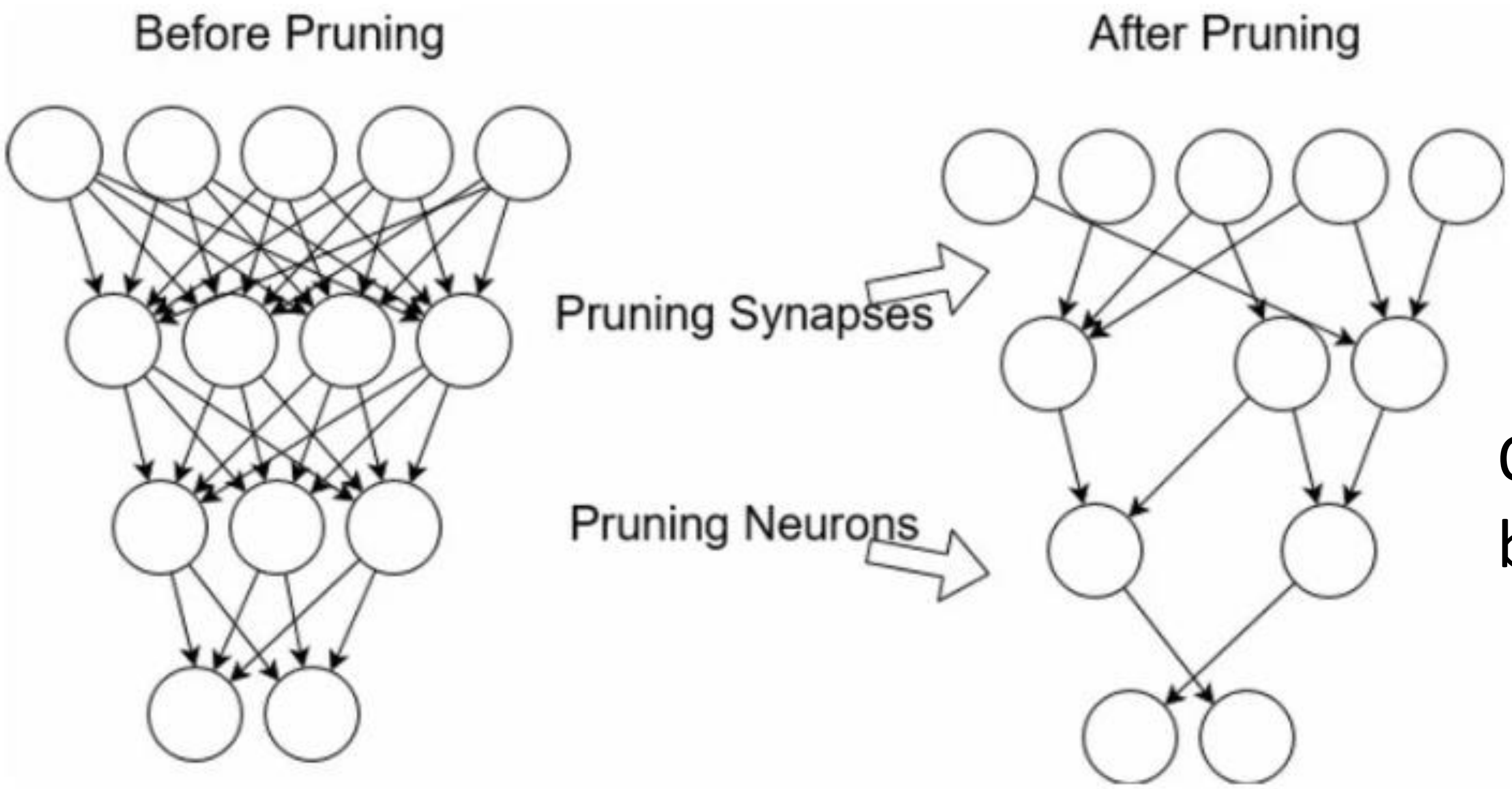


Often remove:
- individual weights
  - e.g., zero low values
- structures
  - e.g., layers

Latter is structural, like a surgery!

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

Pruning: How Much?



Often ~30-50% parameters, based on prior findings

https://www.datature.io/blog/a-comprehensive-guide-to-neural-network-model-pruning
https://xailient.com/blog/4-popular-model-compression-techniques-explained/

What’s Currently Interesting? e.g.,

(arXiv 2025)

Multi-Cue Adaptive Visual Token Pruning for Large Vision-Language Models

Bozhi Luan[1] Wengang Zhou[1] Hao Feng[1] Zhe Wang[2] Xiaosong Li[2] Houqiang Li[1]
[1] University of Science and Technology of China, [2] Huawei Technologies
{bzluan,haof}@mail.ustc.edu.cn, {zhwg,lihq}@ustc.edu.cn, {wangzhe226,lixiaosong20}@huawei.com

Combines spatial information and token similarity to remove redundant tokens

Today’s Topics

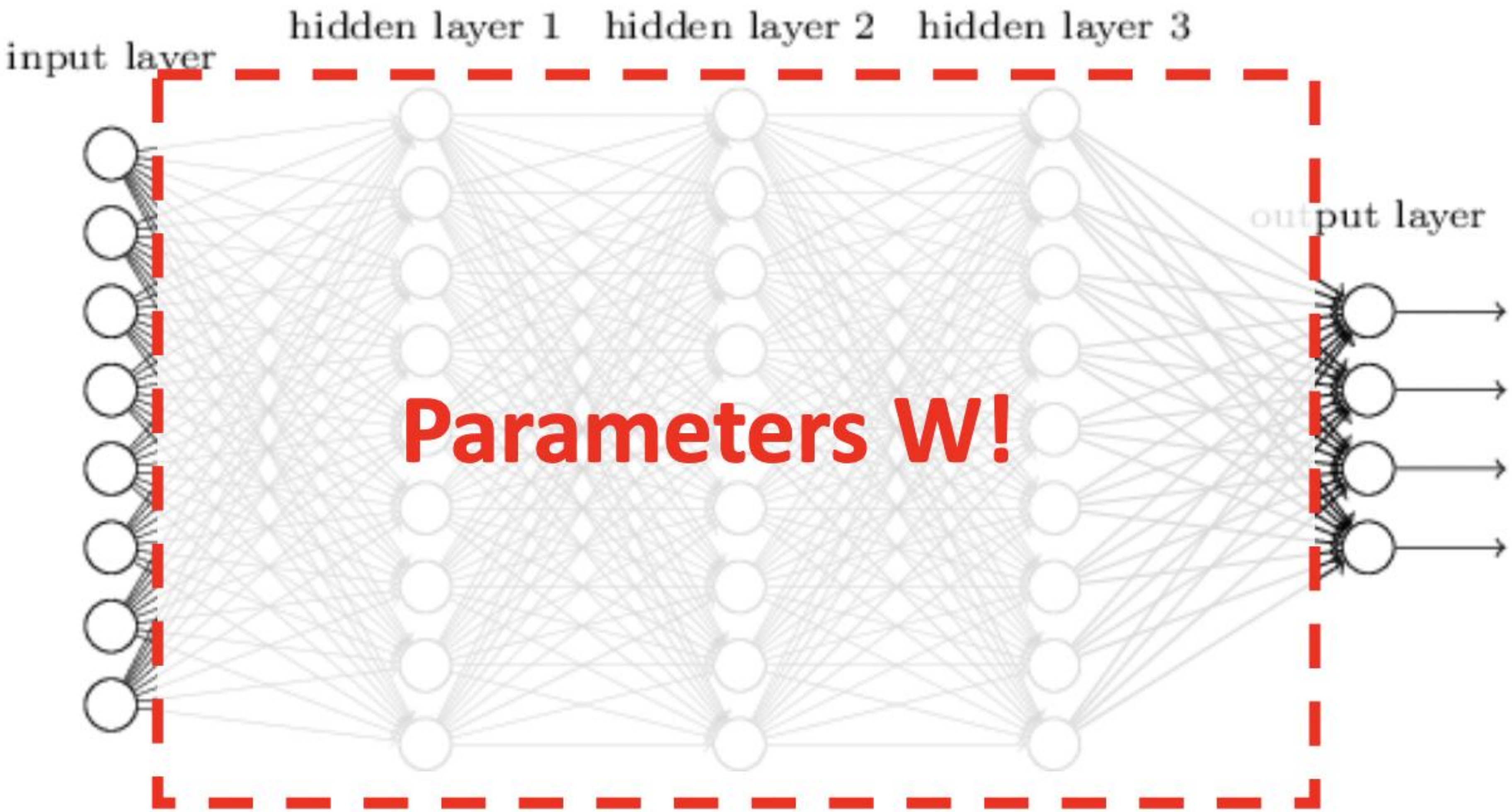
- Motivation
- Pruning
- Knowledge distillation
- Final project report: Overleaf tutorial

Intuition

A student learns from a knowledgeable teacher

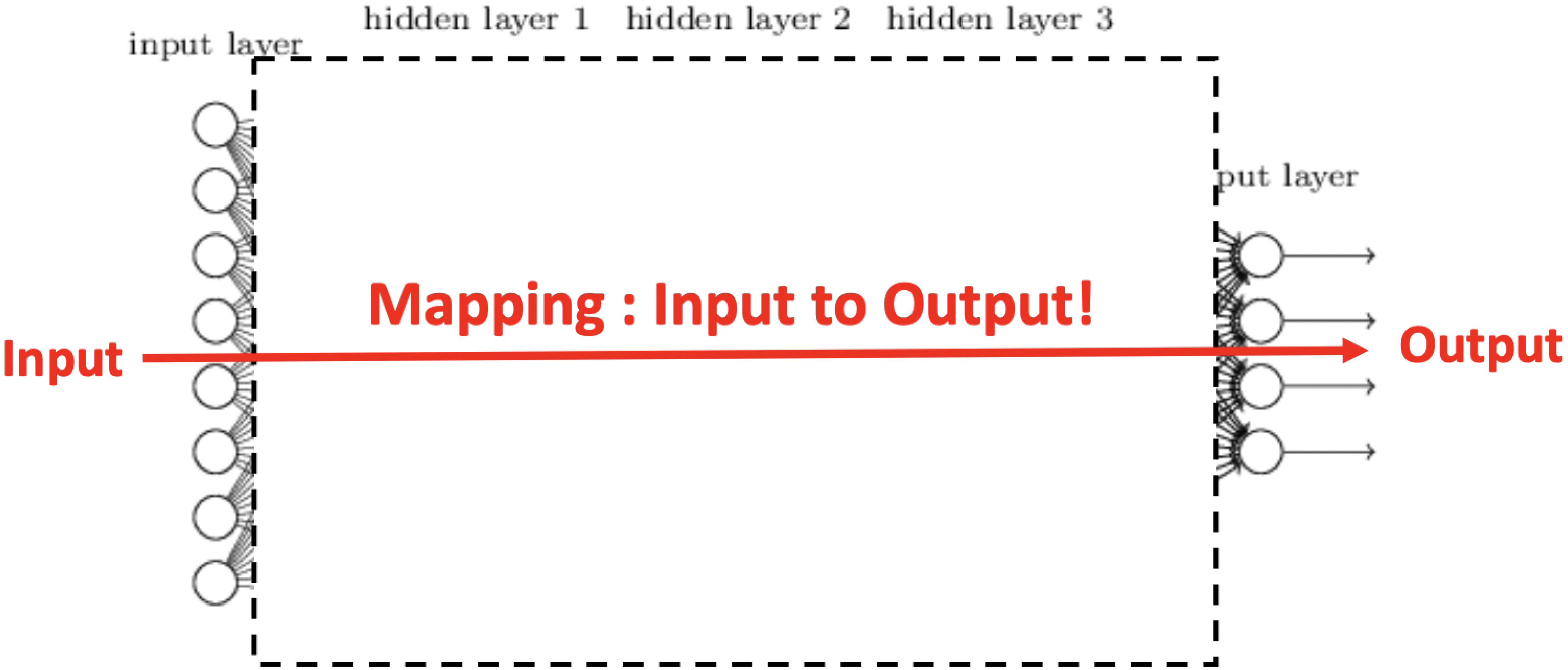
https://www.waterford.org/education/teacher-student-relationships/

Key Question: What is Knowledge?



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

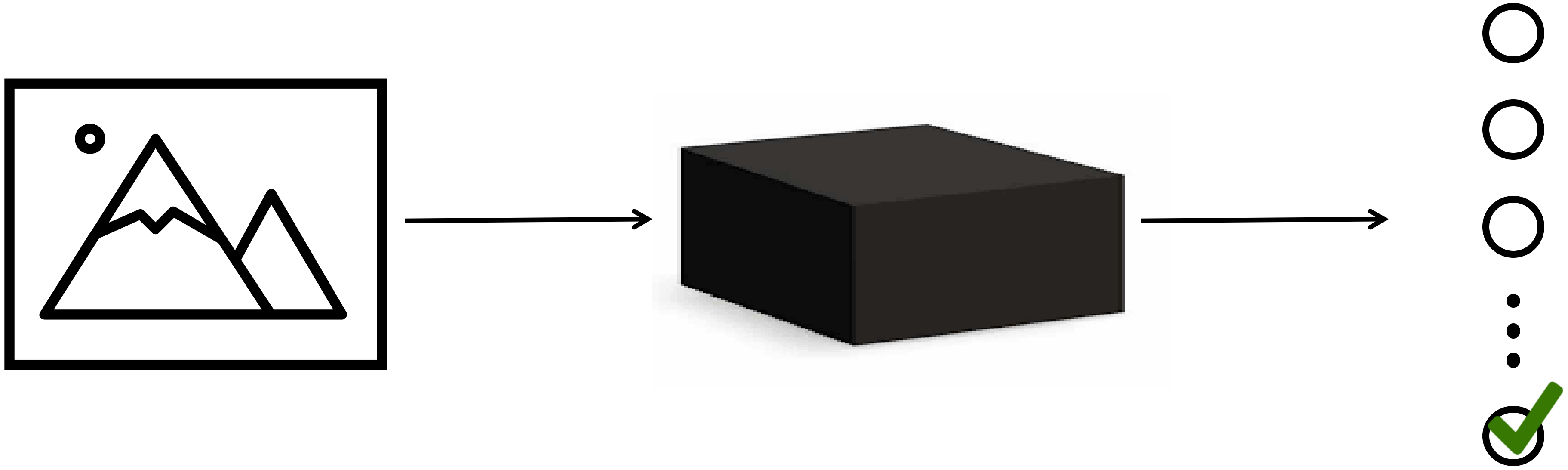
Knowledge Is: Input to Output Mapping



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

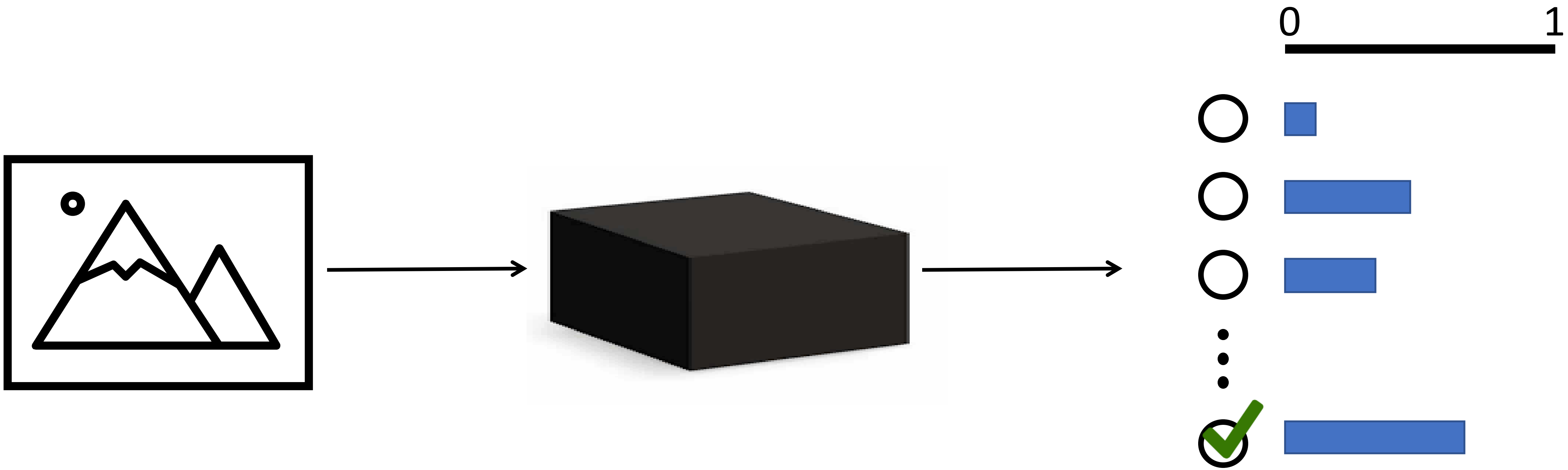
Knowledge Is: Input to Output Mapping

Target mapping: ground truth (1-hot vector)



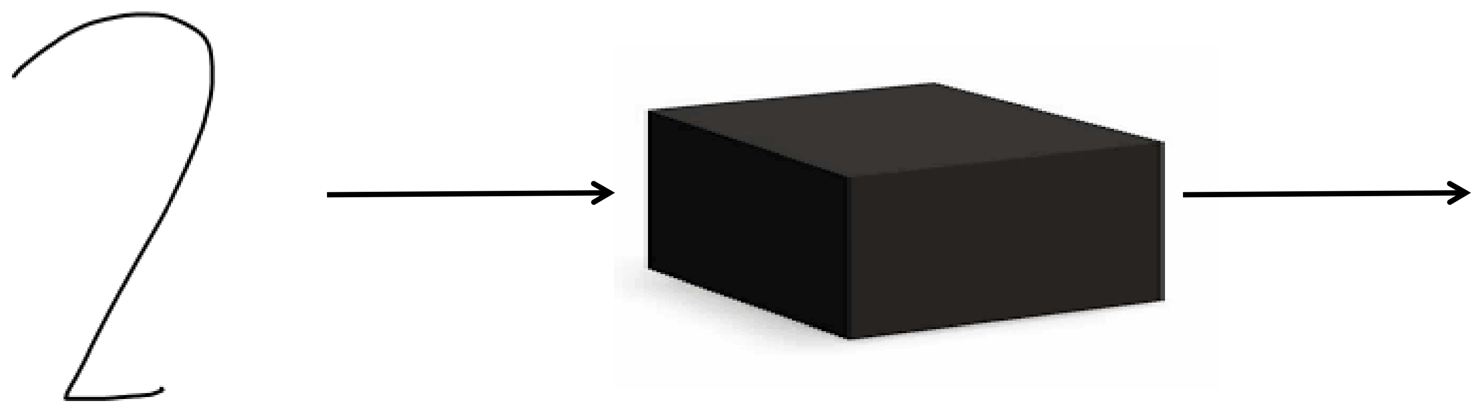
Knowledge Is: Input to Output Mapping

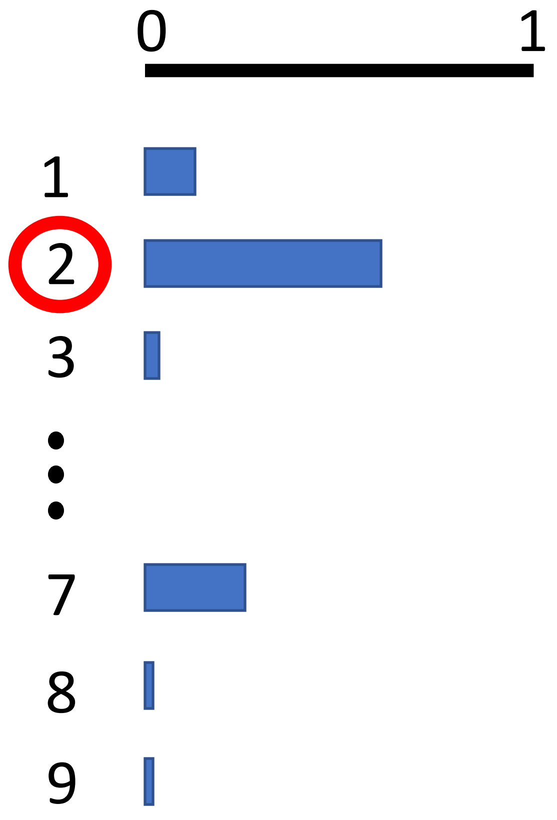
Target mapping: probability distribution from a model offers
further insights into similarities and differences of categories



Knowledge Is: Input to Output Mapping

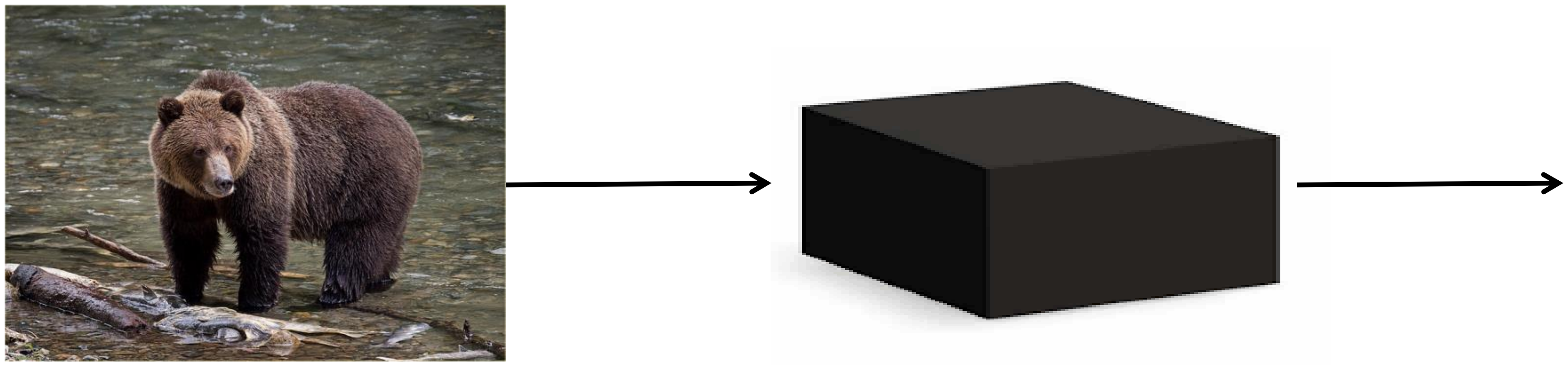
Target mapping: probability distribution from a model offers
further insights into similarities and differences of categories
- Attempts to identify ground truth category
- Also, shares that 2 has similar characteristics to 7 and 1

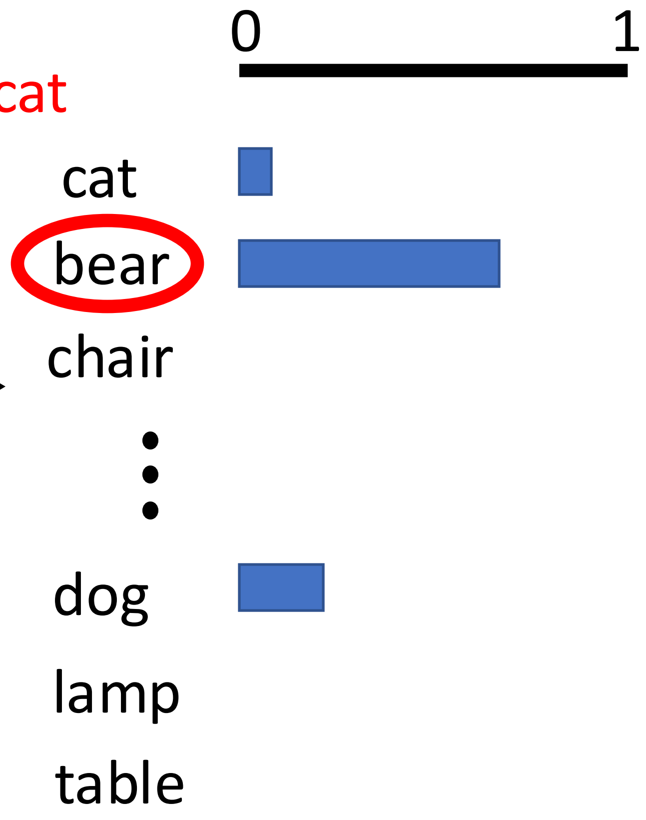




Knowledge Is: Input to Output Mapping

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





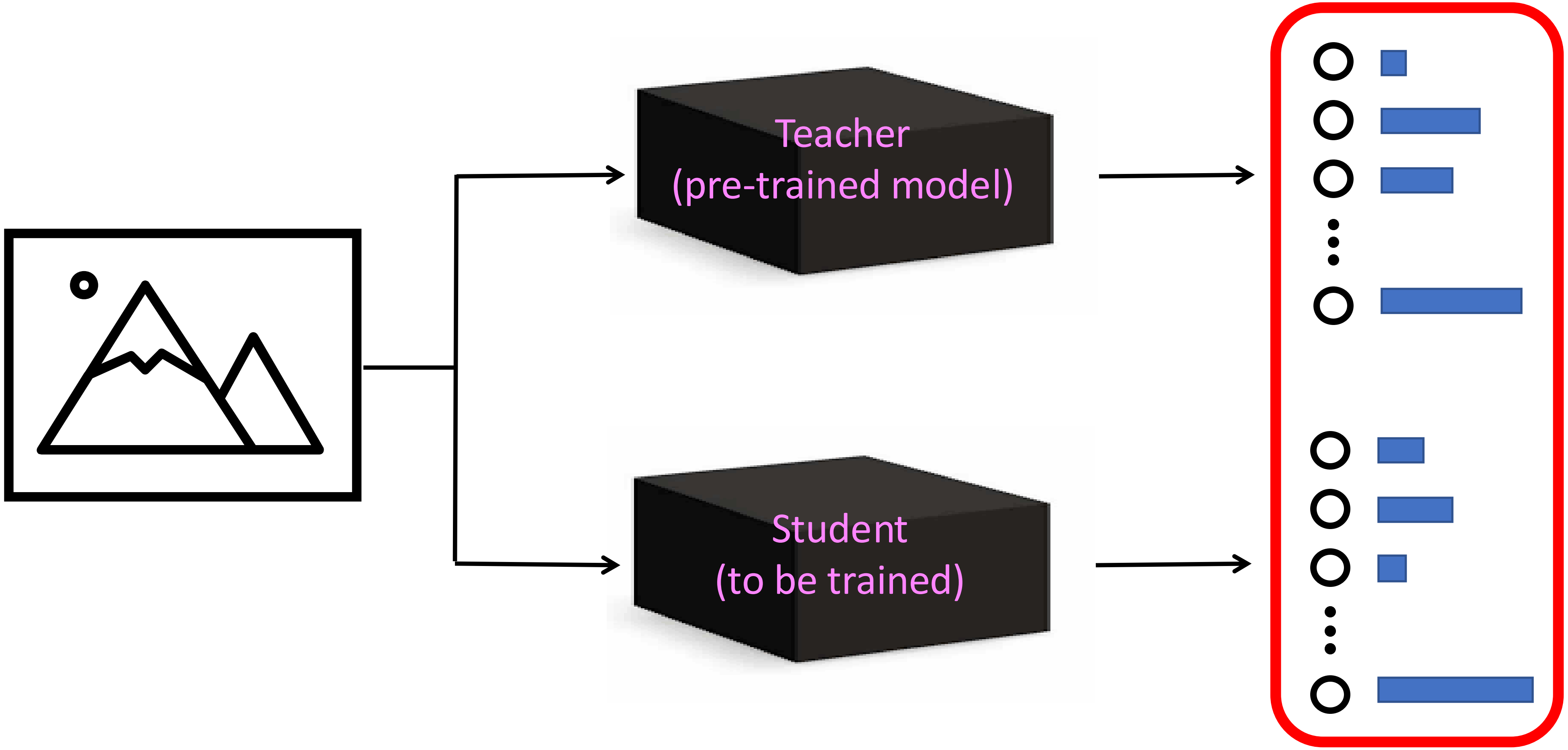
Knowledge Is: Input to Output Mapping

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

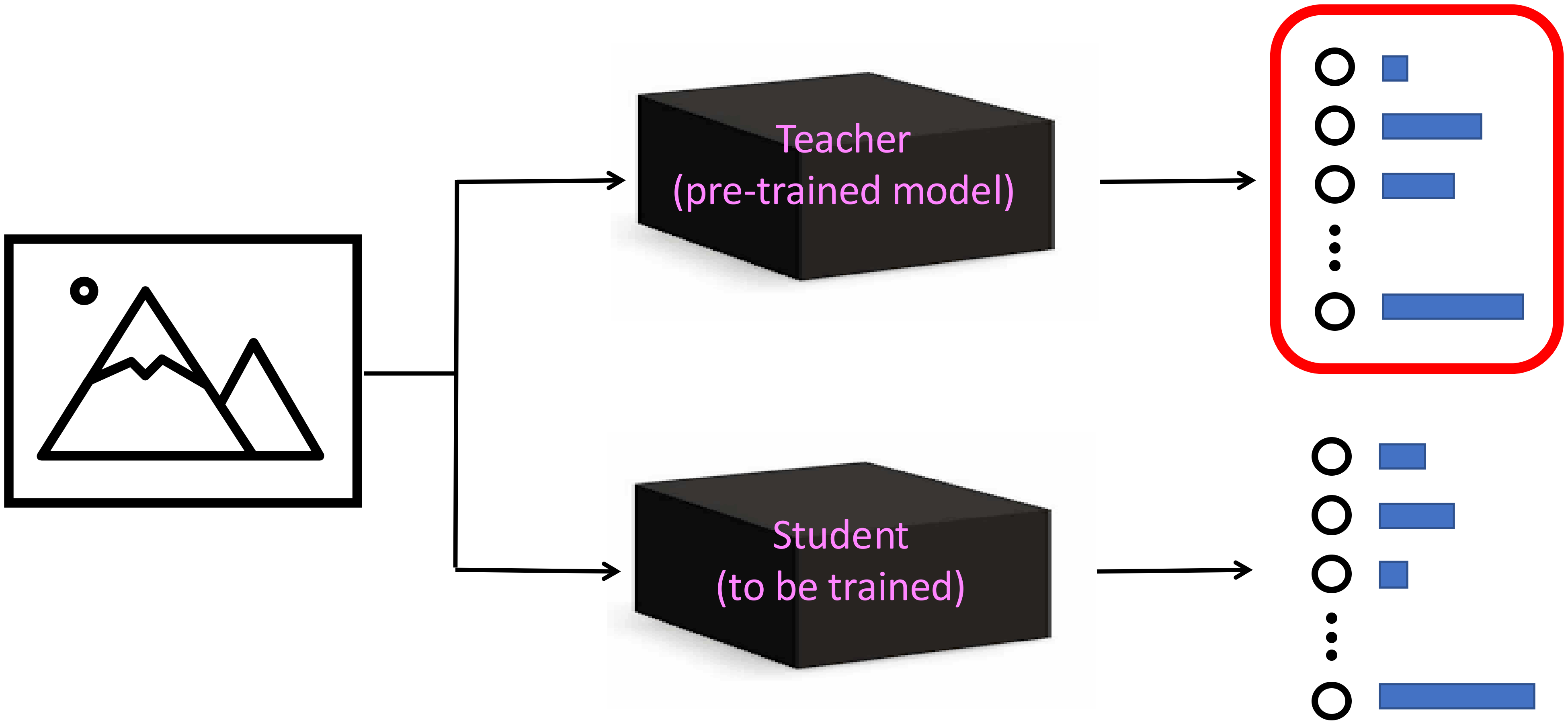
Hinton, Vinyals, and Dean. Distilling the knowledge in a neural network. *arXiv* 2015.

Knowledge Distillation: Teach Student the “Dark Knowledge” of Teacher



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

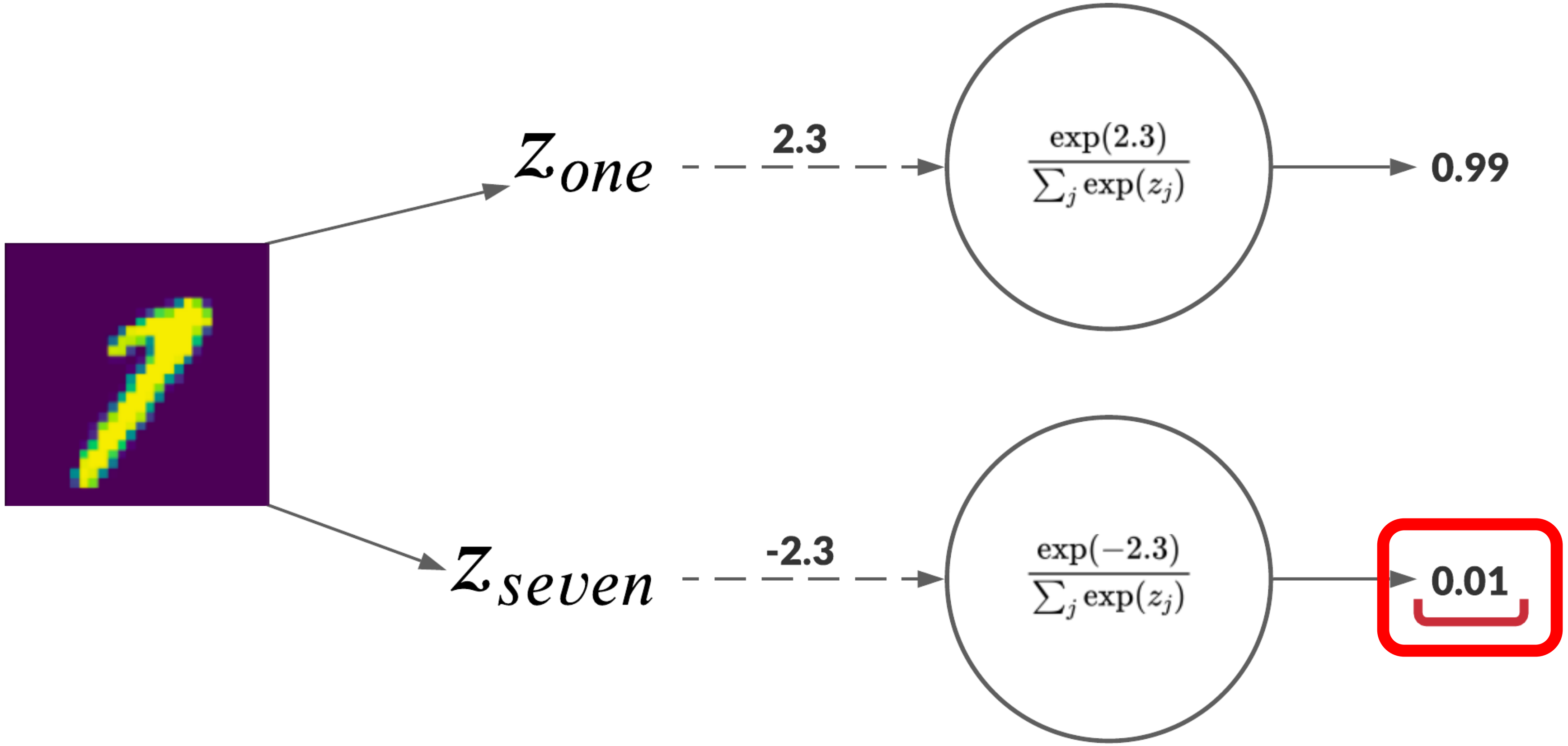
Knowledge Distillation: Teach Student the “Dark Knowledge” of Teacher



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

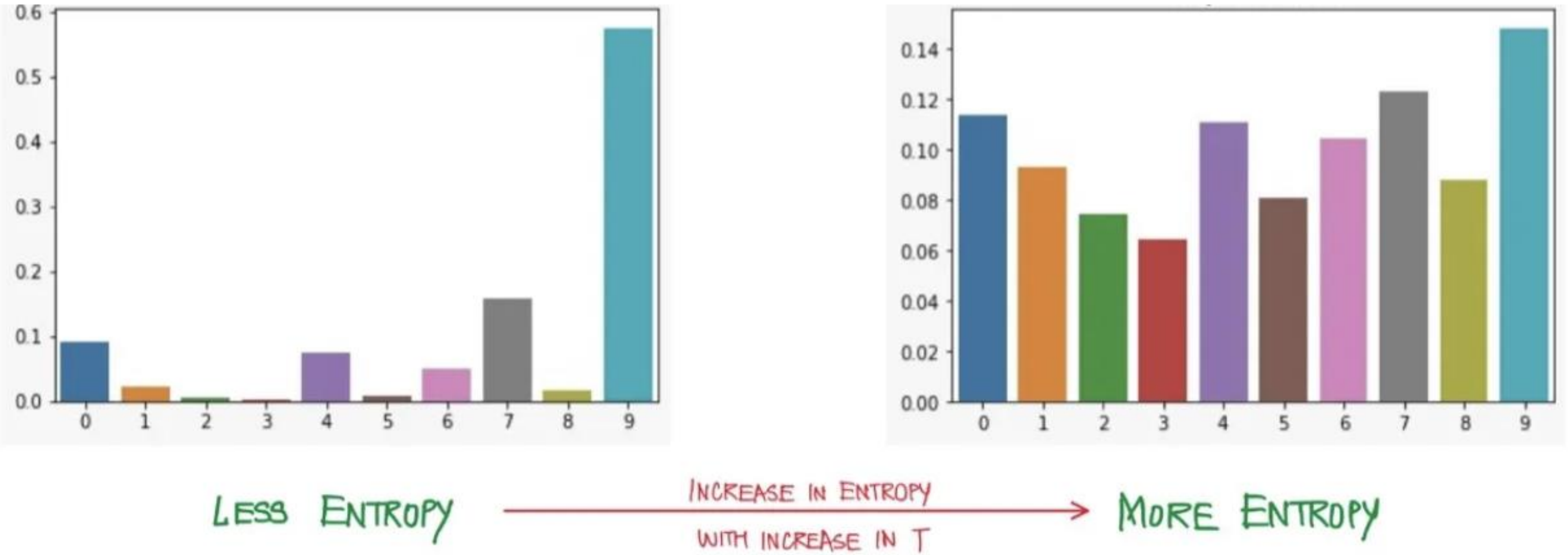
$$\sigma(\mathbf{z})_i = \frac{\exp(z_i / \boxed{T})}{\sum_j \exp(z_j / \boxed{T})}$$

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

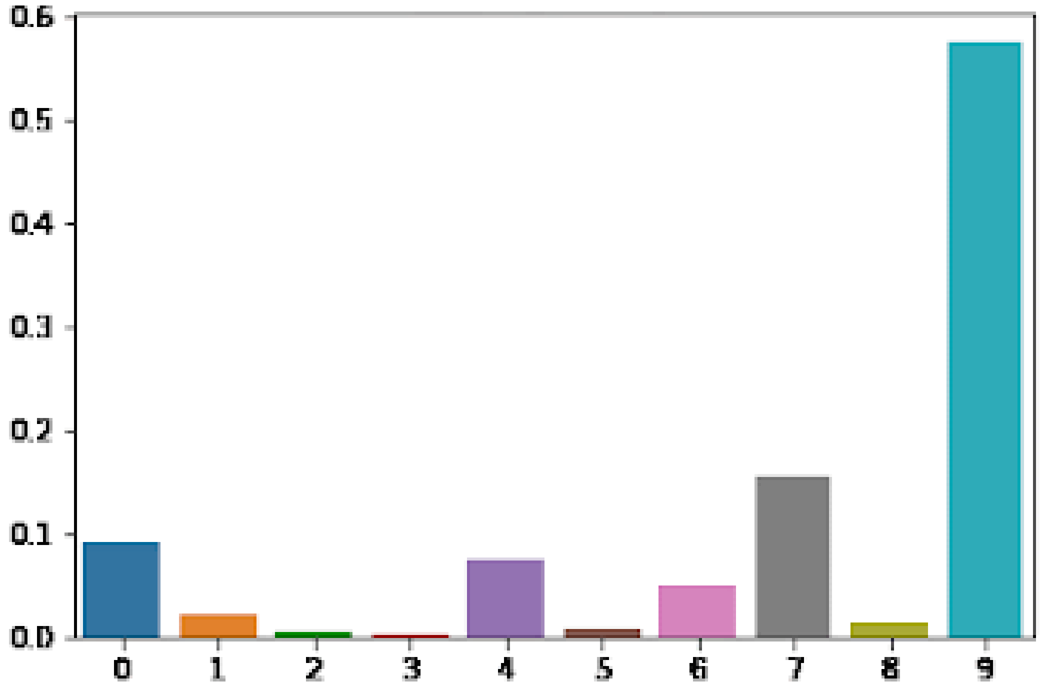
Generalized Softmax: converts scores into a probability distribution summing to 1, with temperature



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

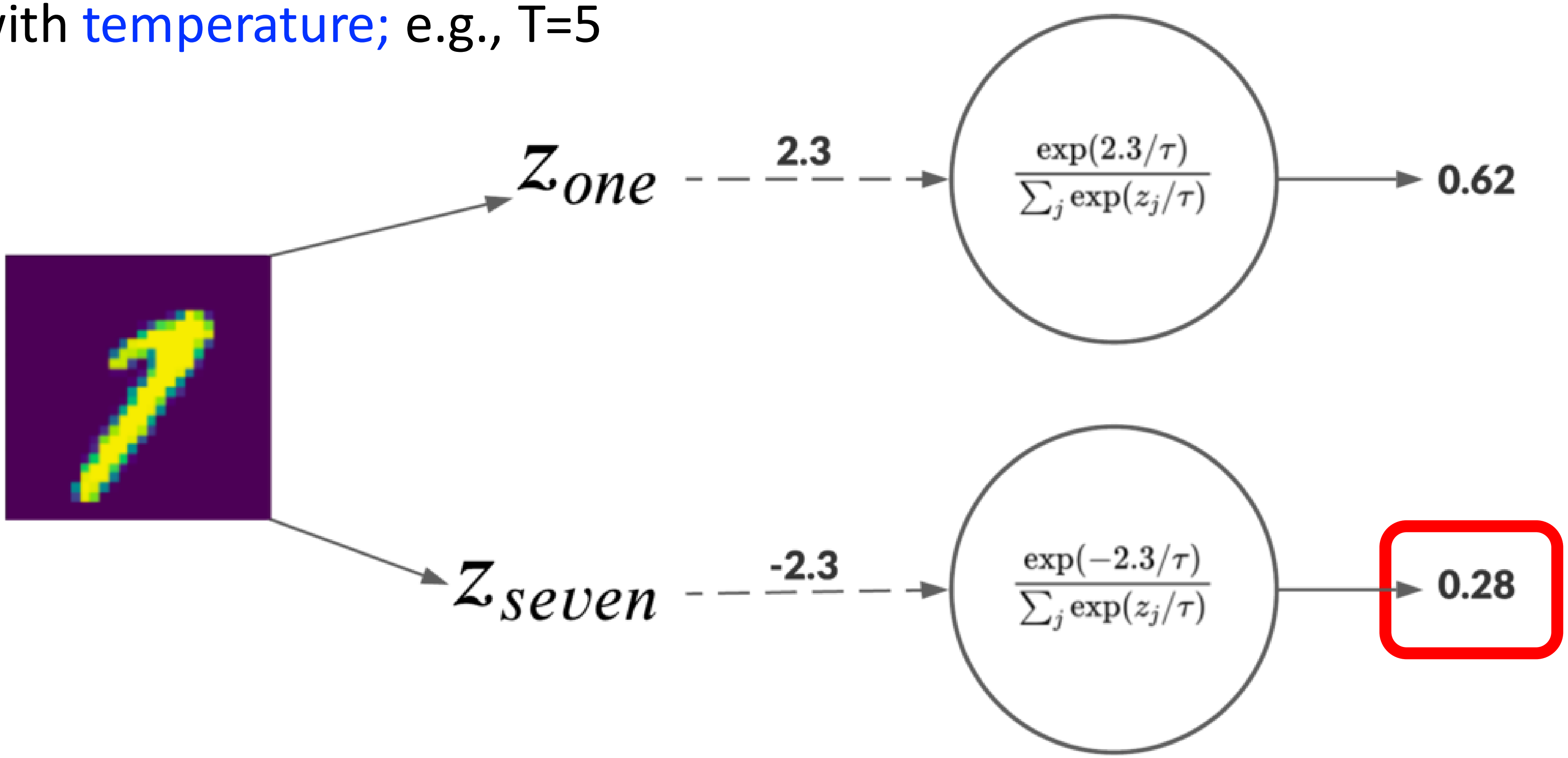


Larger T values provides greater information about which categories are similar to the teacher's predicted category

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

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

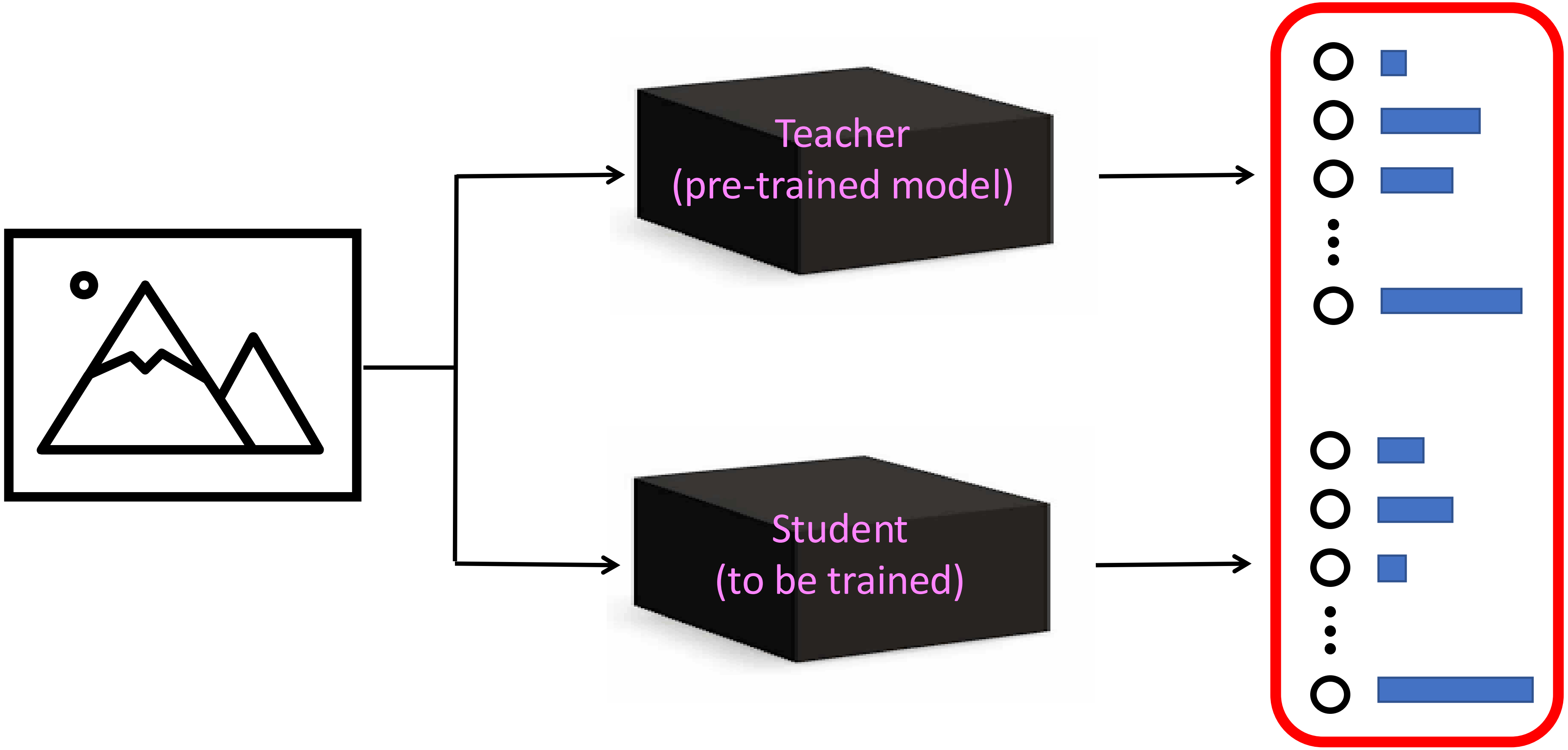
Generalized Softmax: converts scores into a probability distribution summing to 1, with temperature; e.g., T=5



Knowledge of this relationship enhanced from 0.01

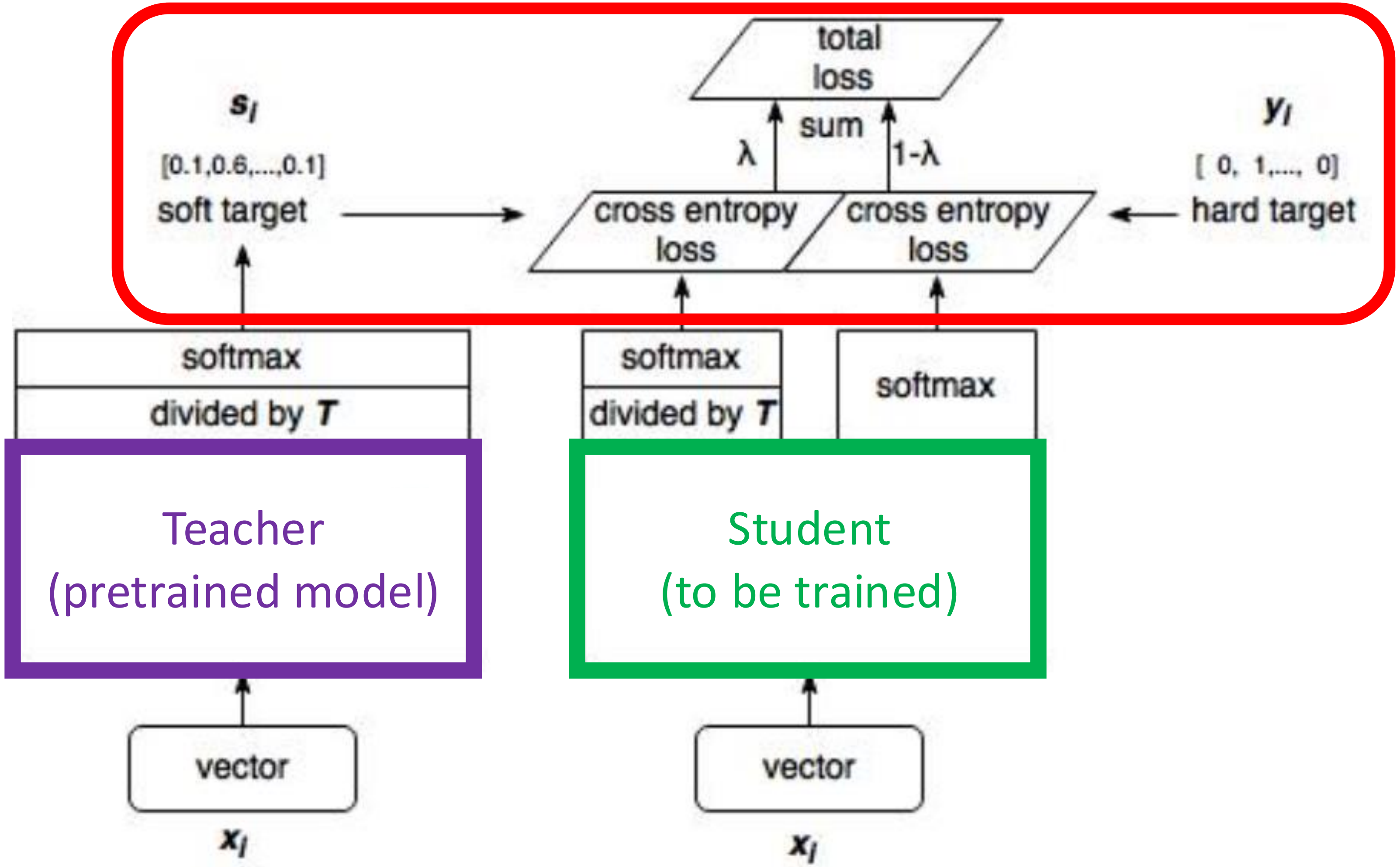
https://wandb.ai/authors/knowledge-distillation/reports/Distilling-Knowledge-in-Neural-Networks--VmlldzoyMjkxODk

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



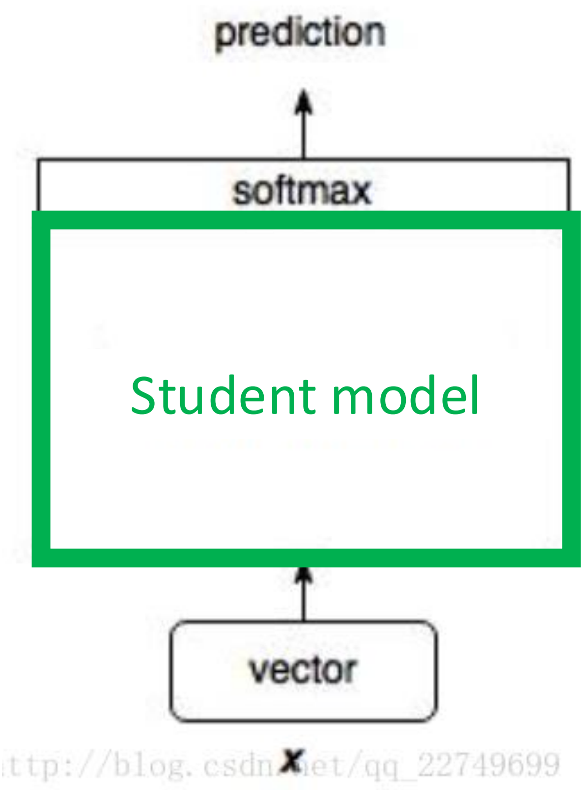
"Distillation" loss computed to bring student's distribution closer to that of the teacher, using the generalized softmax equation

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

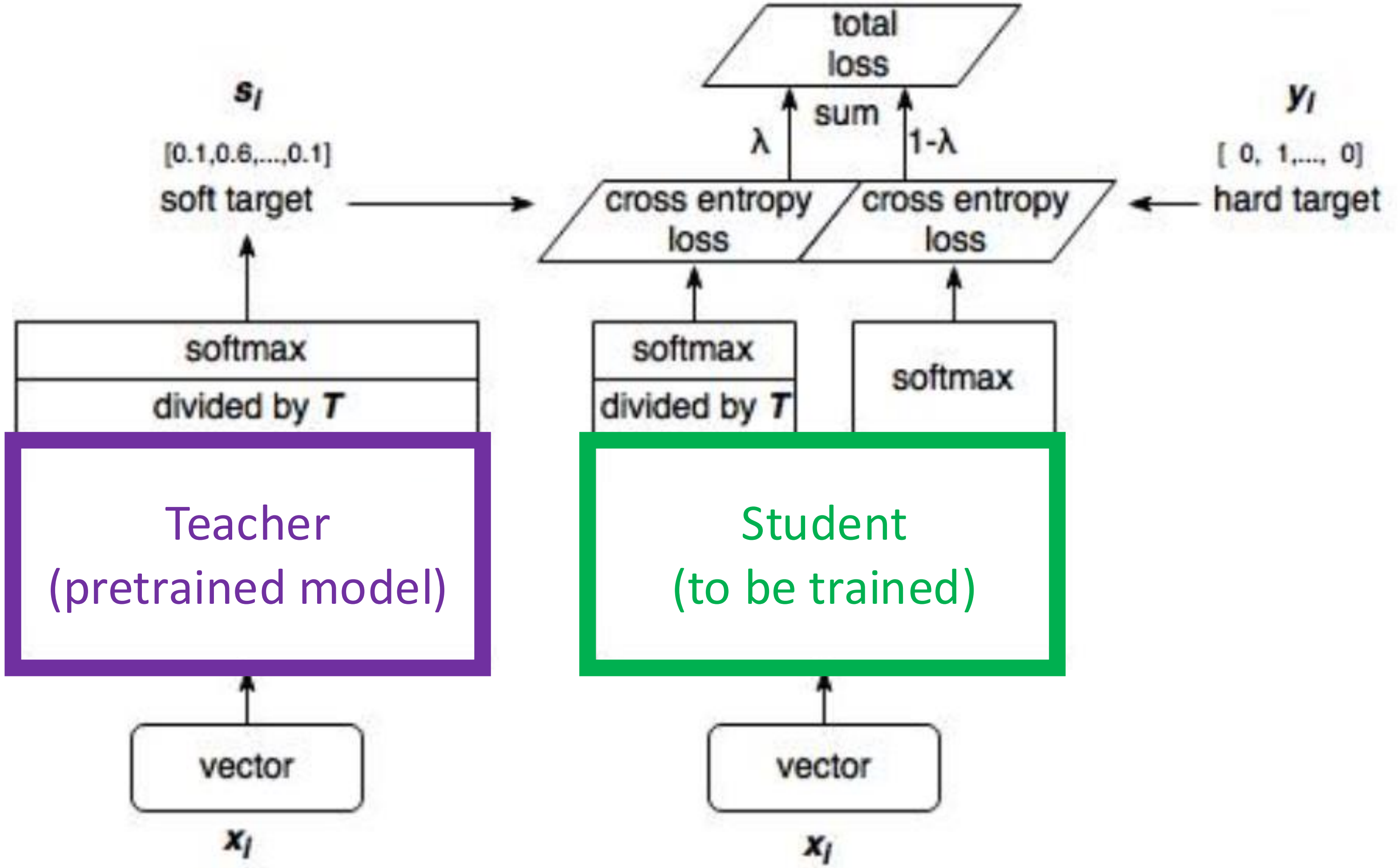


Total loss computed during training is a weighted sum of the conventional cross entropy loss and the "distillation loss"

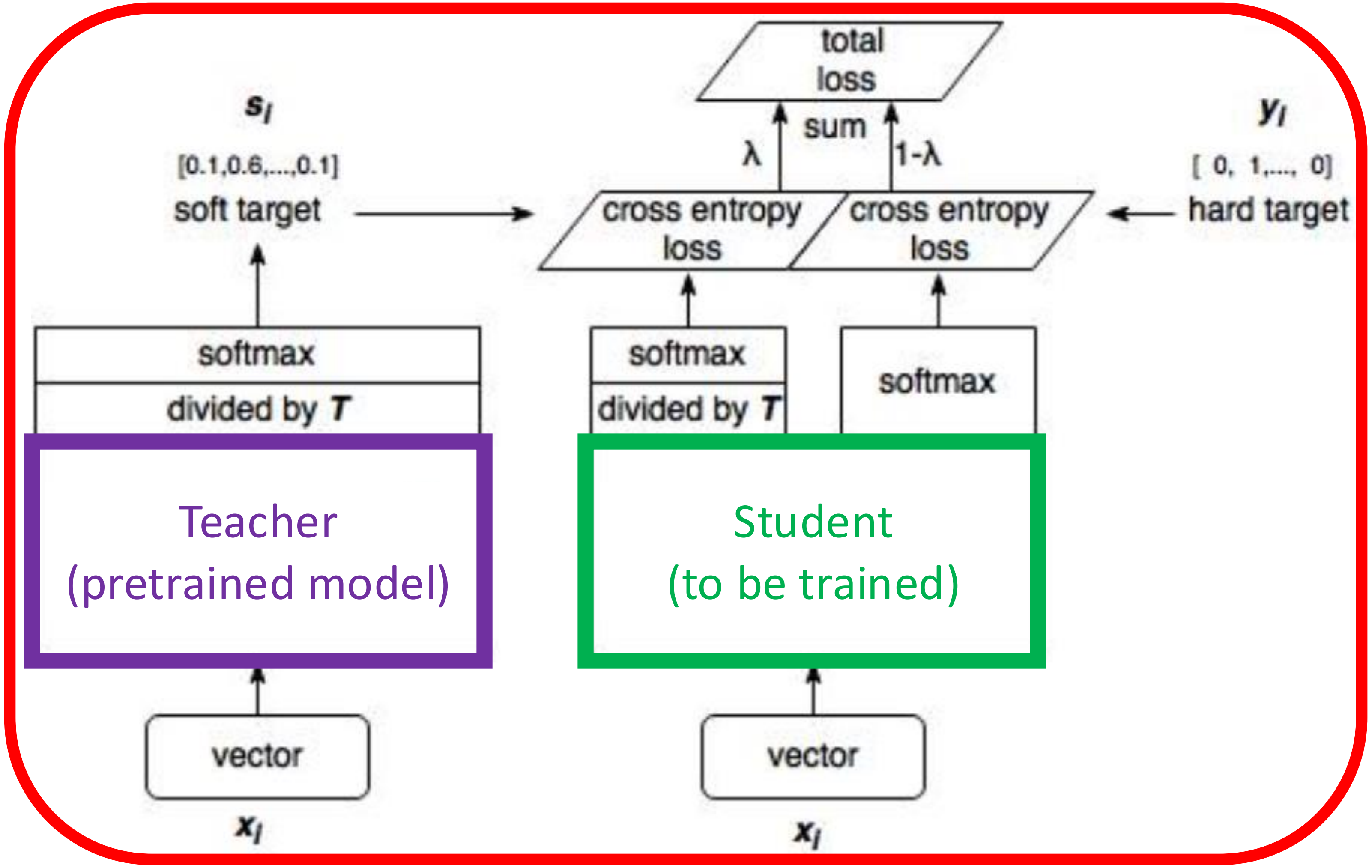
Knowledge Distillation: At Test Time



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



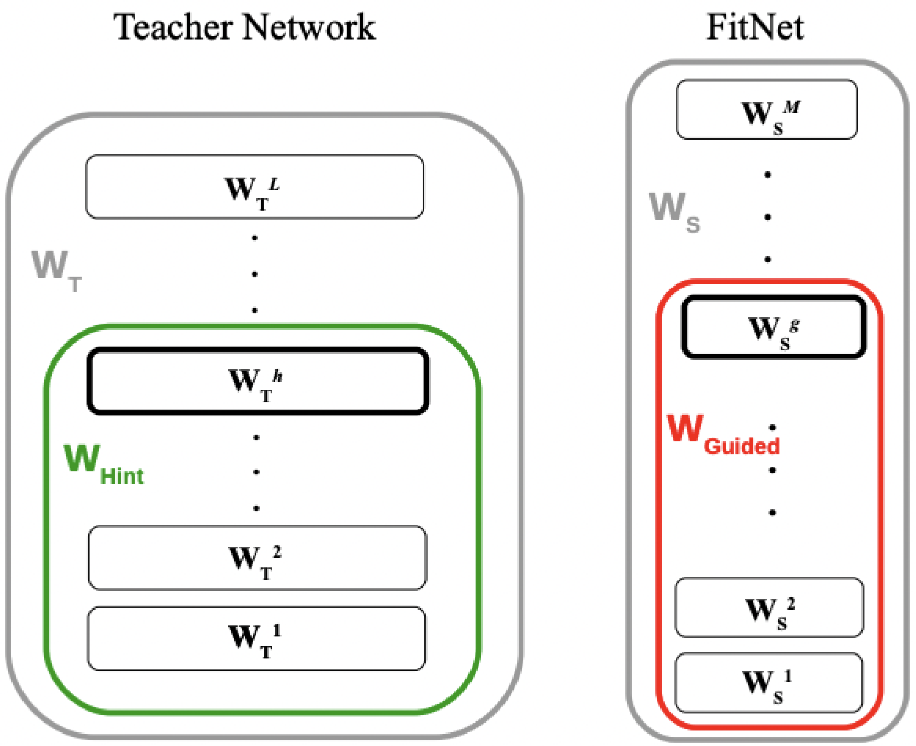
Arguably, Any Neural Network Student Could Learn from Any Neural Network Teacher



Knowledge distillation is a type of transfer learning

Knowledge Distillation Enhancement: Hints

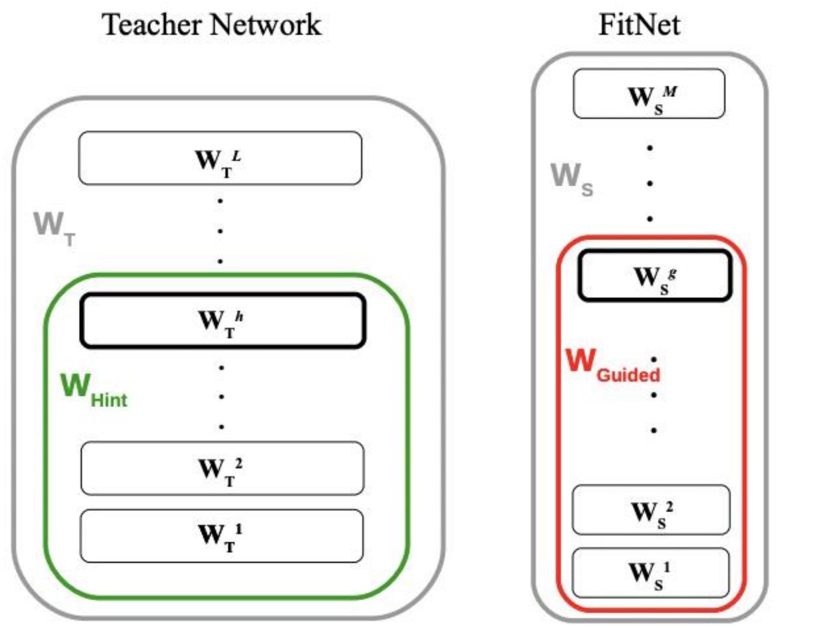
Encourage student (FitNet) to mimic the teacher's feature responses; e.g., output of guided layer should match the output of hint layer



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

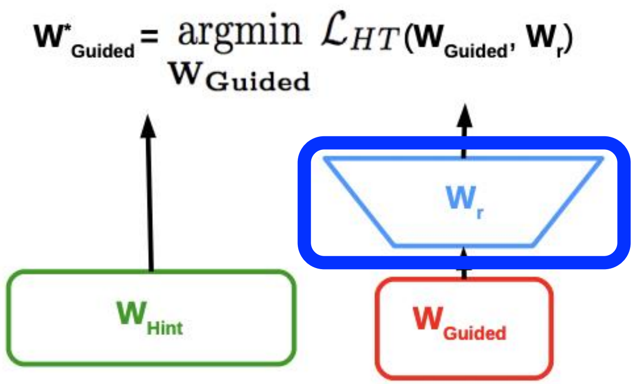
Knowledge Distillation Enhancement: Hints

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

Training conducted to learn the intermediate feature



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

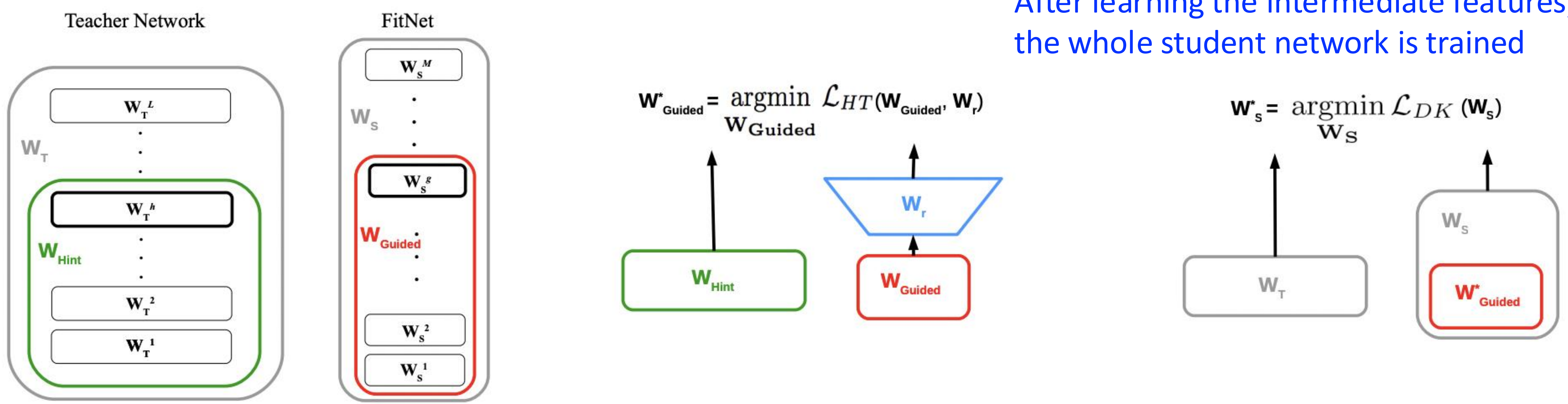
(b) Hints Training

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

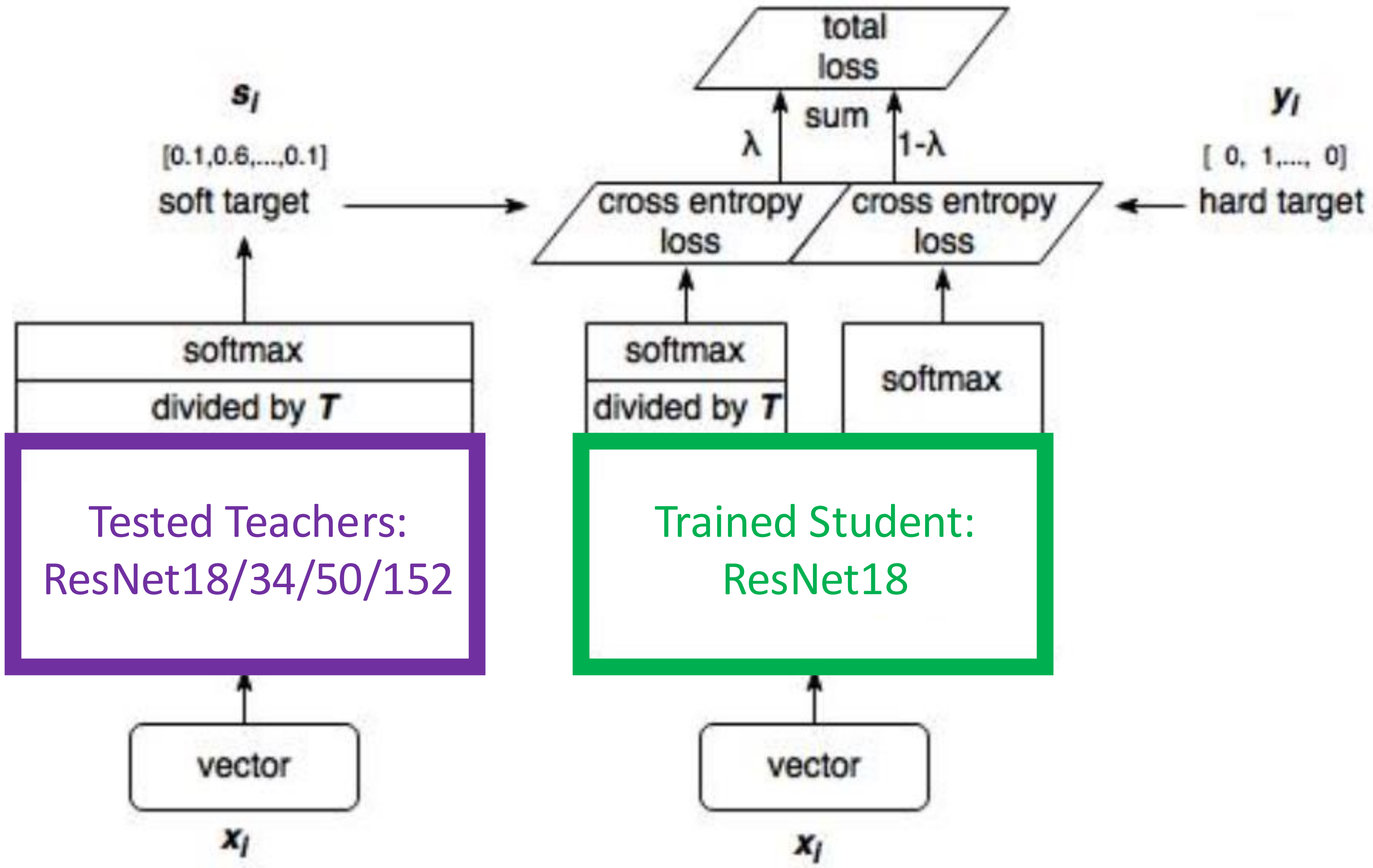
Romero et al. Fitnets: Hints for thin deep nets. ICLR 2015.

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

Deng et al. 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; 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

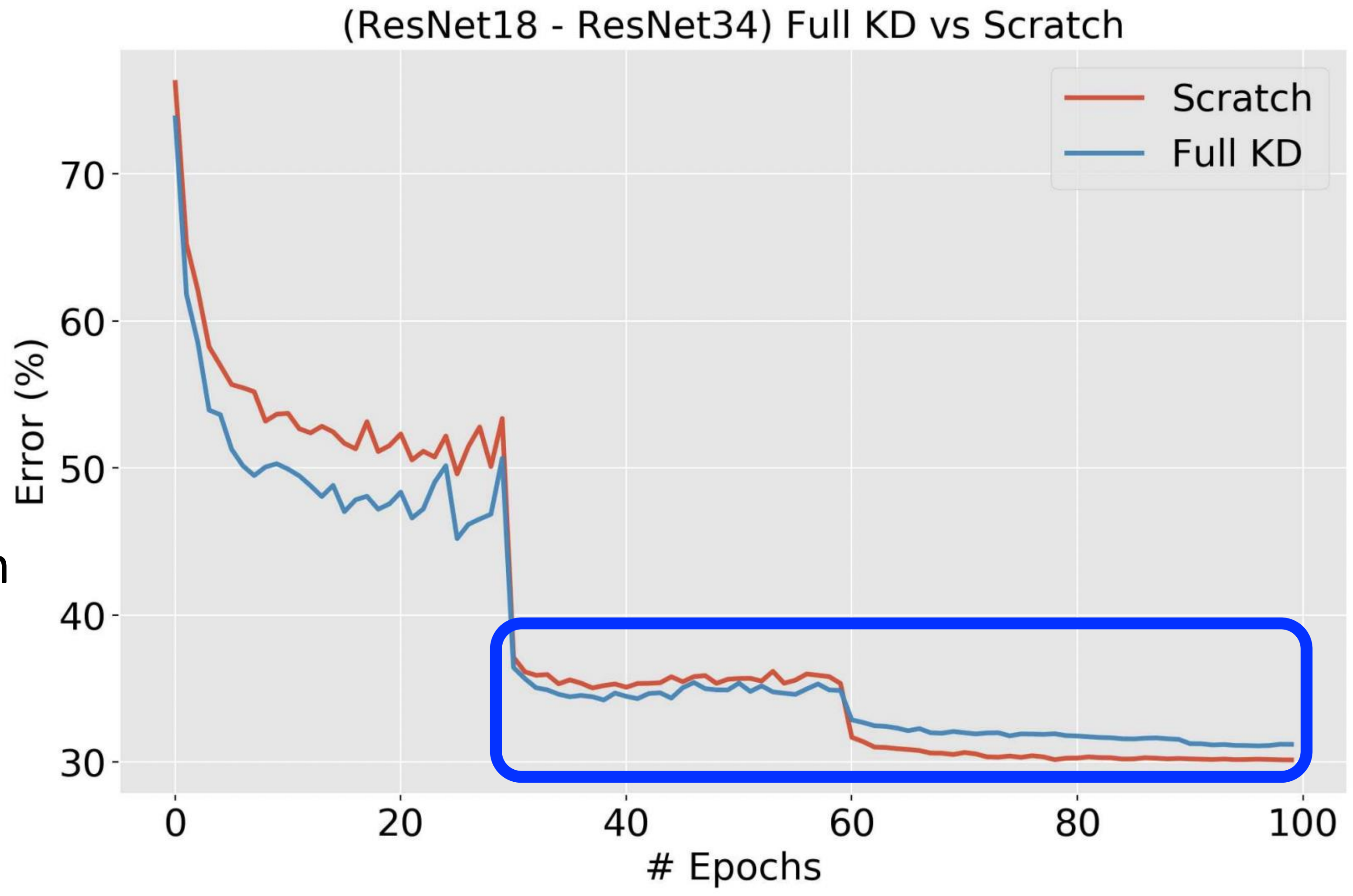
2. Student mimics teacher but the loss function is mismatched from the evaluation metric

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?



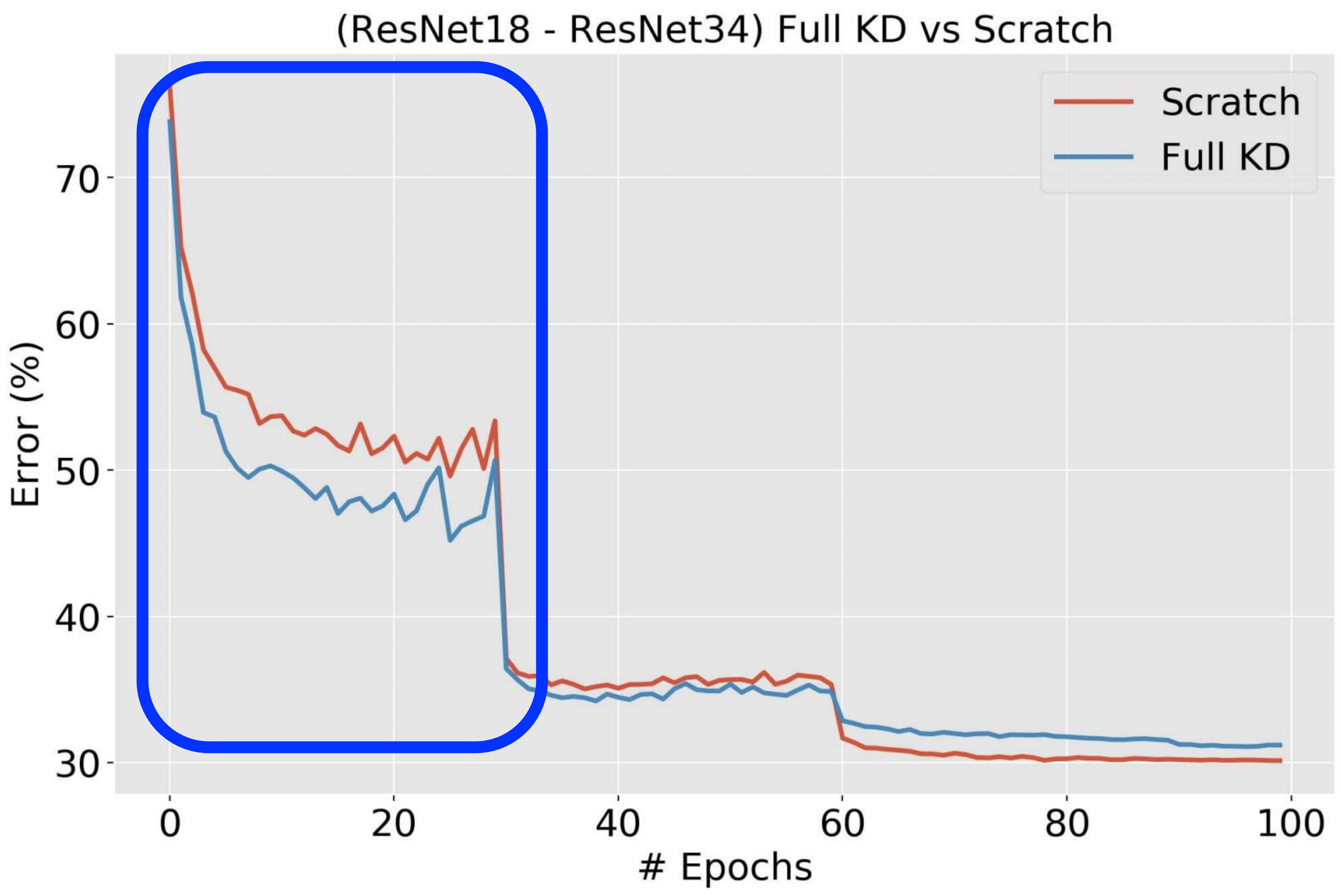
Hypothesis: student underfitting from small capacity and so "minimizing one loss (KD loss) at the expense of the other (cross entropy loss)"

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.,

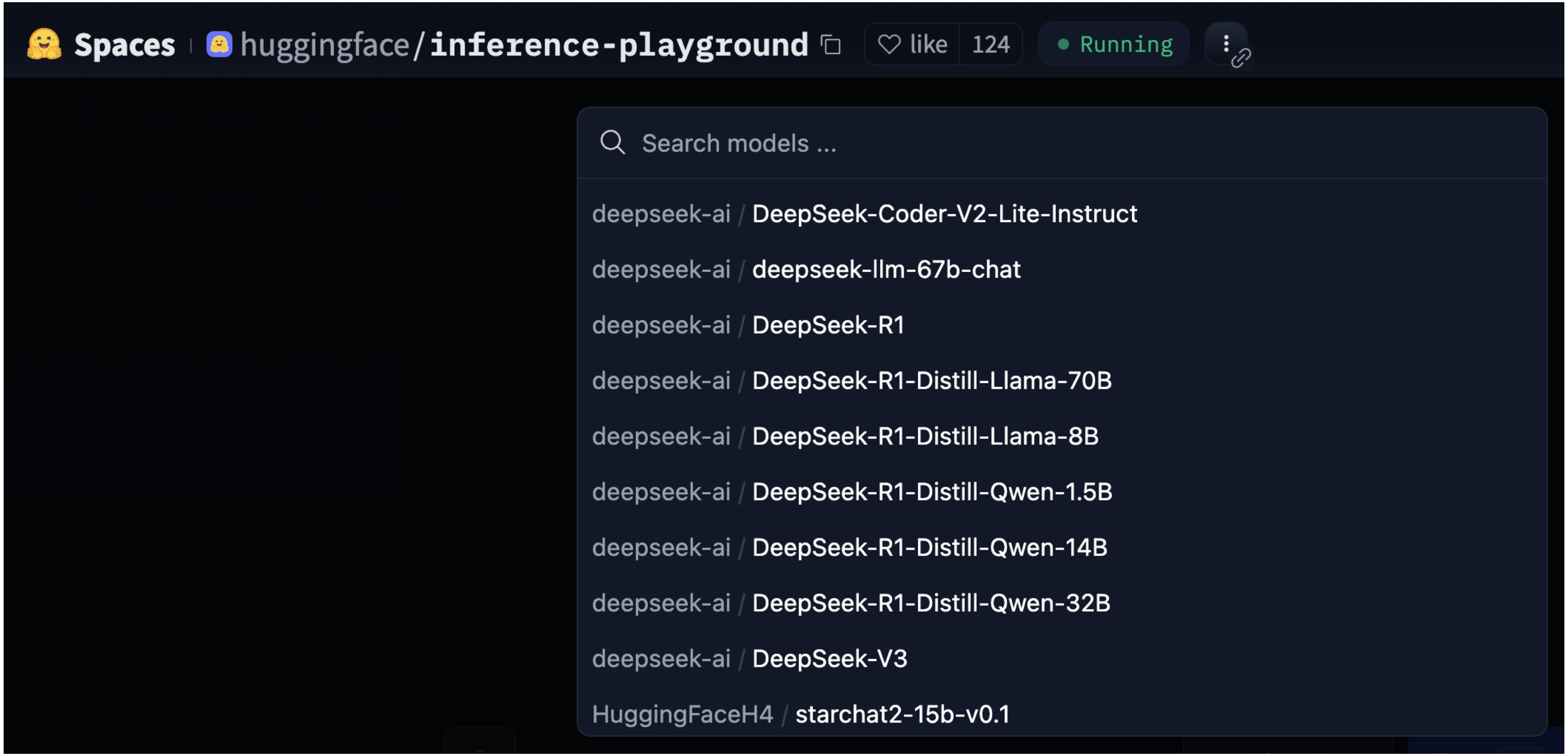
(Neurips 2024)

What Knowledge Gets Distilled in Knowledge Distillation?

Utkarsh Ojha* Yuheng Li* Anirudh Sundara Rajan*

Yingyu Liang Yong Jae Lee

Today’s Status Quo: Multiple Model Sizes; e.g.,



Today’s Topics

- Motivation
- Pruning
- Knowledge distillation
- Final project report: Overleaf tutorial

Project Outline

- Latex in Overleaf to create professional-looking document

- Related work: How to find them?
  - Market search
  - Publication search (e.g., Google Scholar)
  - New AI tools
    - e.g., “Deep Research” from OpenAI, with prompt: https://taaft.notion.site/The-Ultimate-Deep-Research-Prompt-Engineer-1c0ed82cbfd38068aa61ff326133fc24
    - e.g., AI2's paper finder: https://allenai.org/blog/paper-finder

Today’s Topics

- Motivation
- Pruning
- Knowledge distillation
- Final project report: Overleaf tutorial


The End