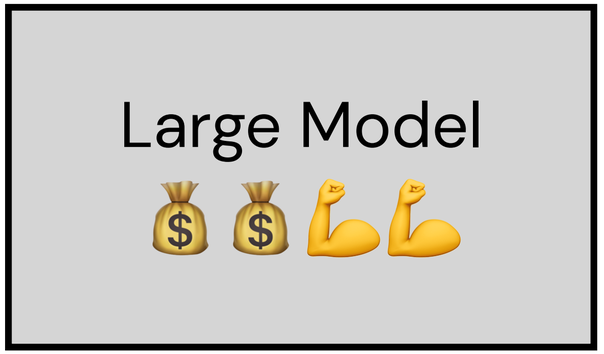
Knowledge Distillation

EECE695D: Efficient ML Systems

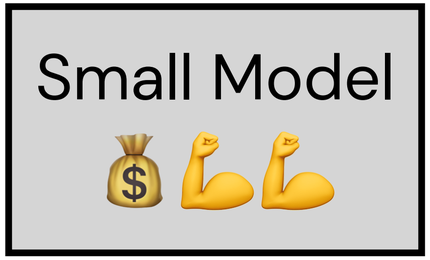
Spring 2025

Recap

- **Problem.** Inference cost
  - Large models are strong but heavy

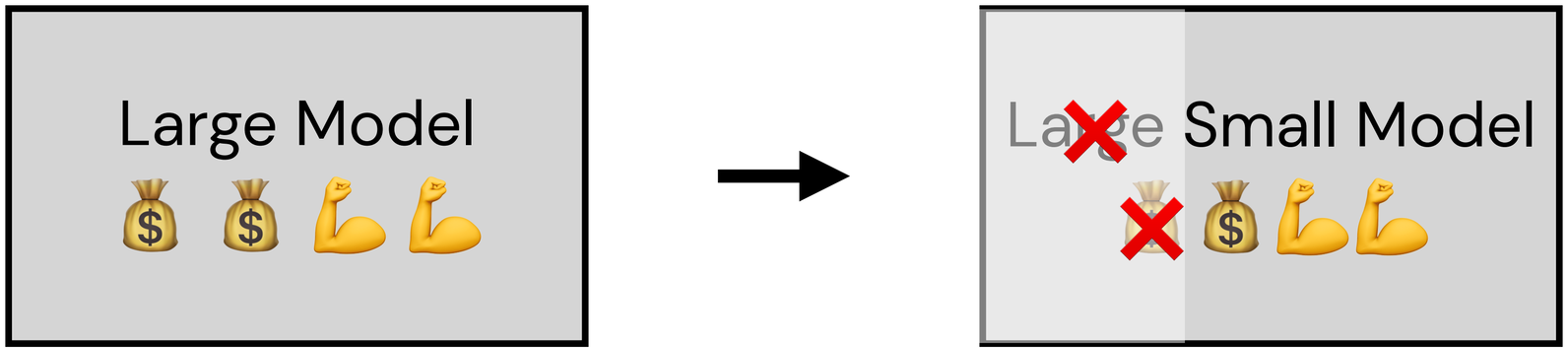


- **Goal.** Lowering the inference cost
  - Make a small but strong model



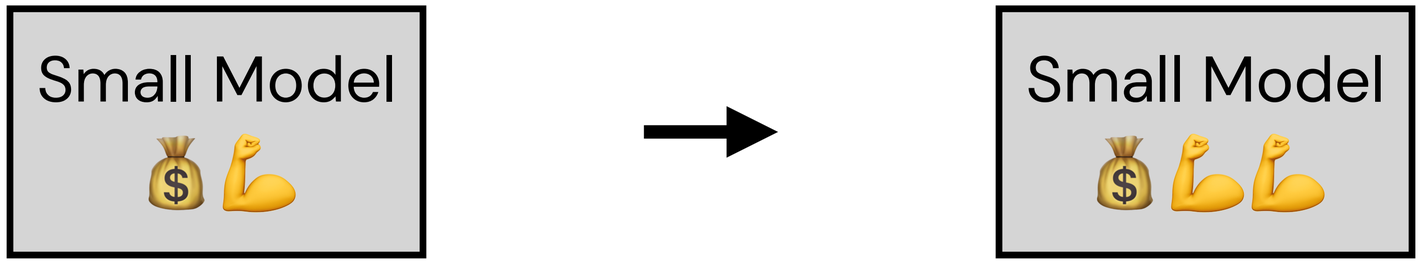
Recap

- **W2-W3.** Reduce the cost of large models
  - Idea. Replace matmuls with lighter options
    - Sparsity
    - Low-Precision



Recap

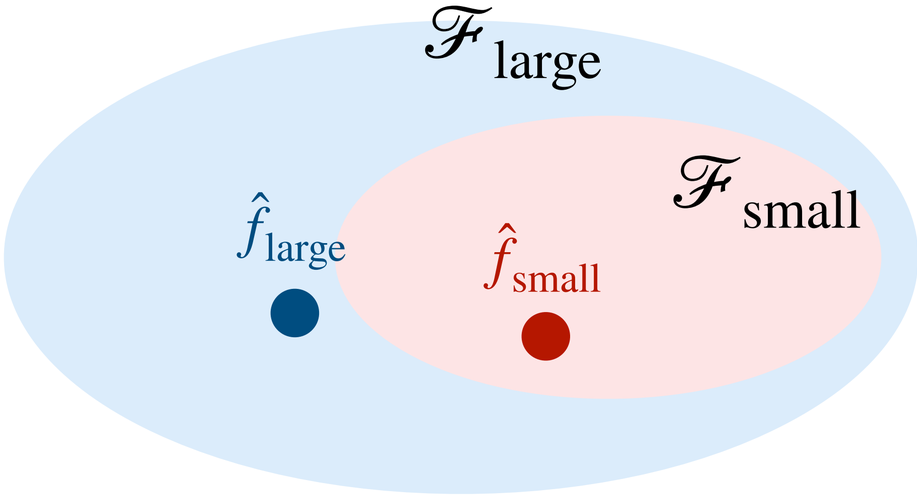
- **Today.** Making small models stronger
    - Knowledge Distillation



Basic idea

Motivation

- **Small model.** $\hat{f}_{\text{small}} = \arg\min_{f \in \mathcal{F}_{\text{small}}} \hat{L}(f)$
- **Large model.** $\hat{f}_{\text{large}} = \arg\min_{f \in \mathcal{F}_{\text{large}}} \hat{L}(f)$



- **Question.** Why do $\hat{f}_{\text{small}}$ work much poorer than $\hat{f}_{\text{large}}$?

  (A) $\mathcal{F}_{\text{small}}$ does not have a good optimum

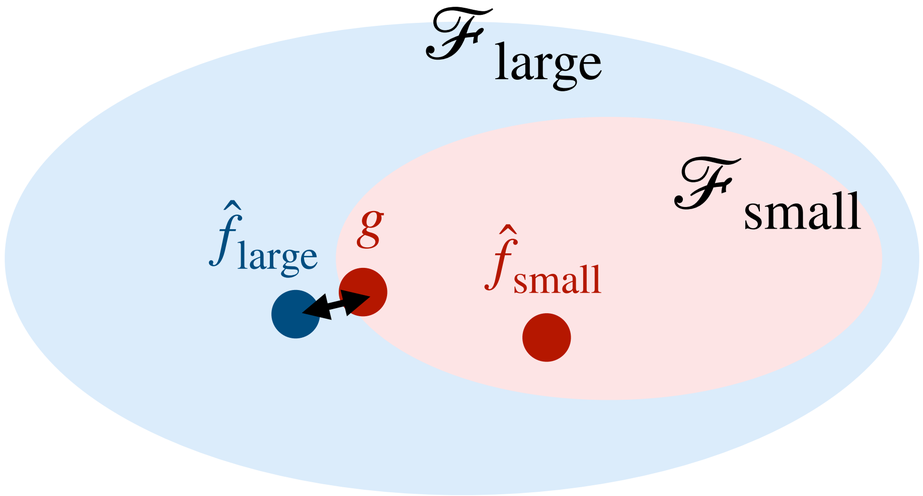
  (B) Good optimum exists, but is difficult to find

Motivation

(A) $\mathcal{F}_{\text{small}}$ does not have a good optimum $\Rightarrow$ **False**

- **Theoretically.** Few weights are enough to approximate any function (link)
  - Thus, it is likely that there exists some g such that

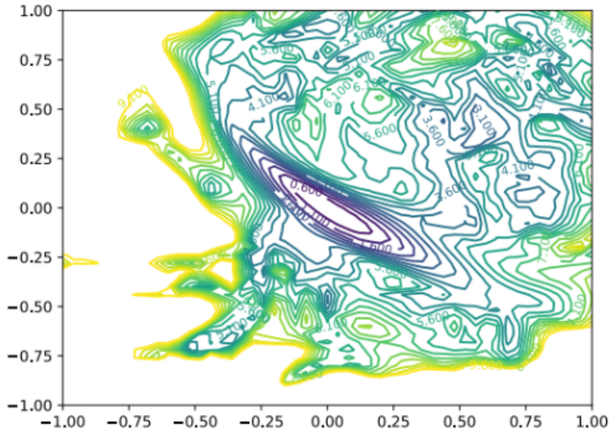
$$g \approx \hat{f}_{\text{large}}, \qquad g \in \mathcal{F}_{\text{small}}$$



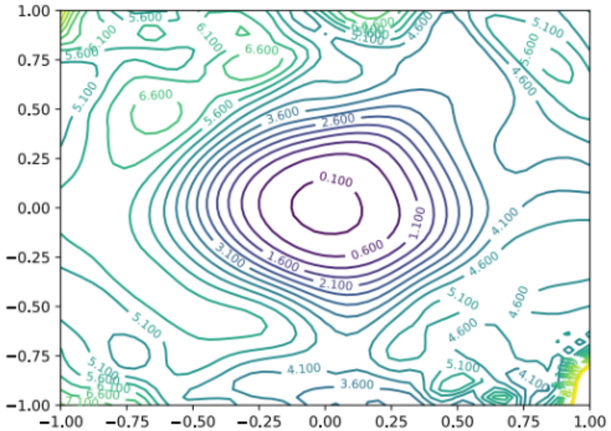
Motivation

(B) Good optimum exists, but is difficult to find ⇒ **True**

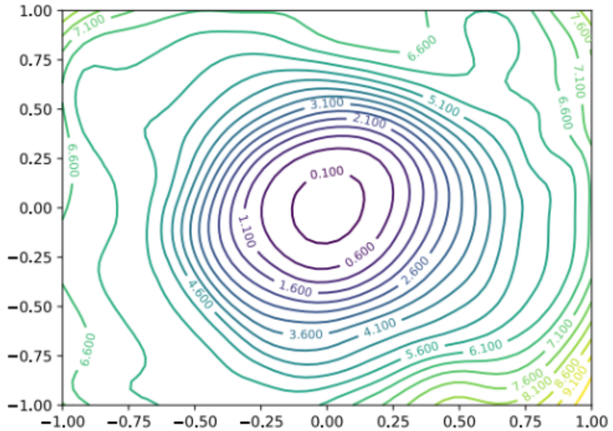
- **Empirically.** GD on small nets suffer from local minima
  - Example. Loss landscape of ResNet-like models (narrow -> wide)



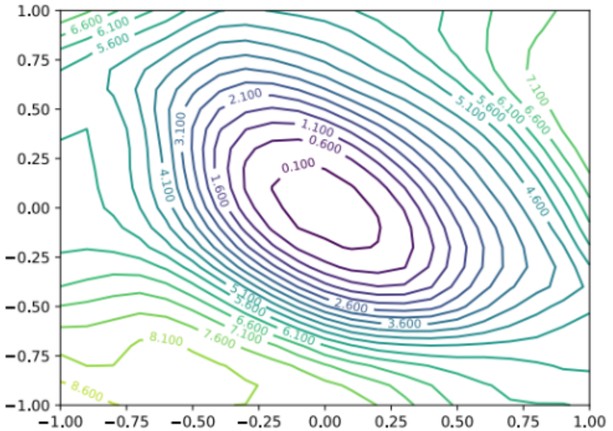
(e) $k = 1, 13.31\%$



(f) $k = 2, 10.26\%$



(g) $k = 4, 9.69\%$



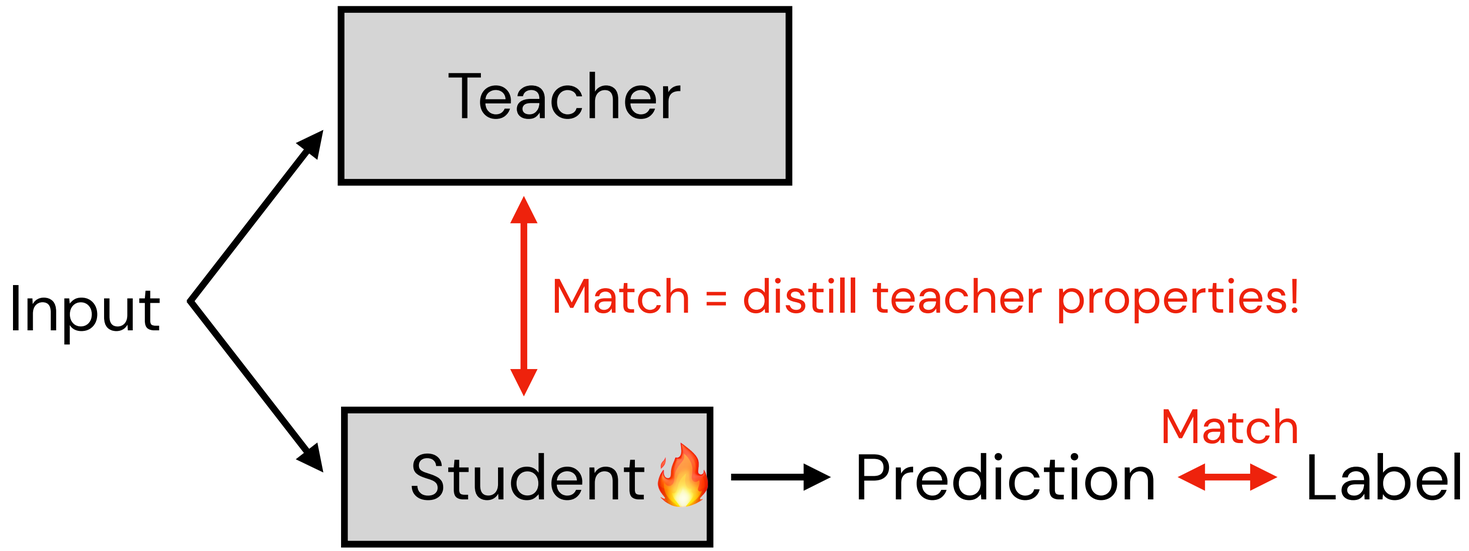
(h) $k = 8, 8.70\%$

Li et al., "Visualizing the loss landscape of neural nets," NeurIPS 2018

Knowledge distillation

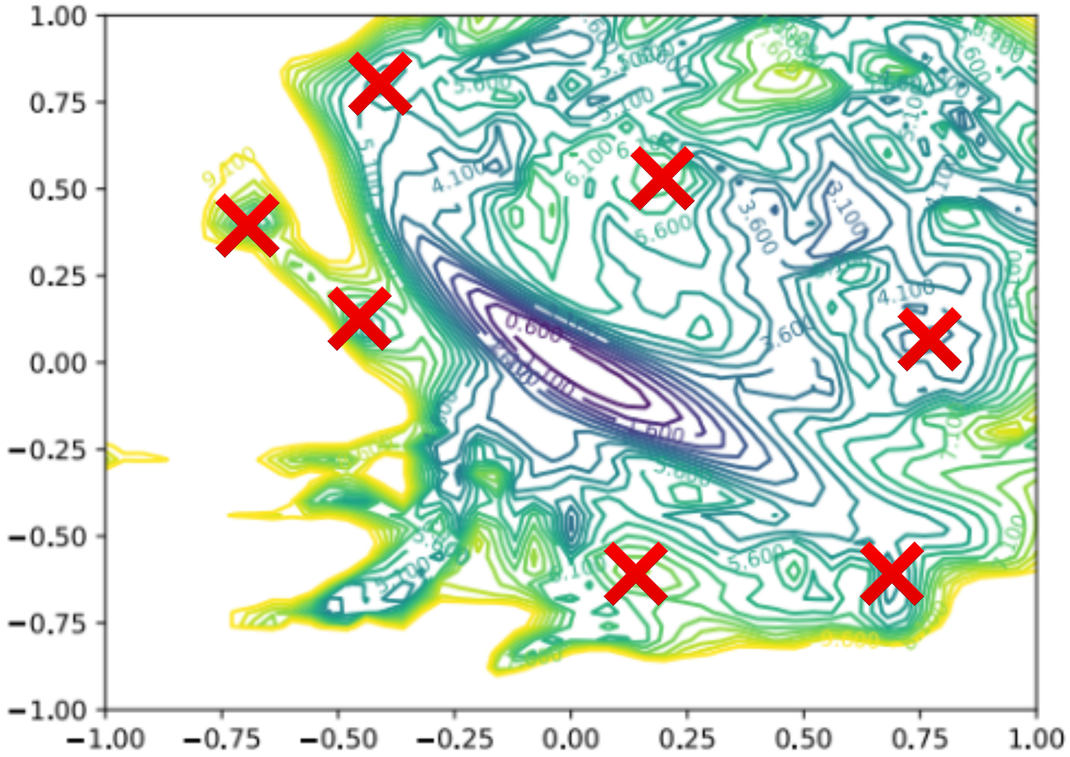
- **Question.** How do we encourage the small model to find a nice function?
- **Idea.** Explicitly regularize the learned function to be functionally close to the pre-trained large model $\hat{f}_{\text{large}}$ (also called "teacher")

$$\min_{f \in \mathcal{F}_{\text{small}}} \hat{L}(f) + \text{dist}(f, \hat{f}_{\text{large}})$$



Knowledge distillation

- This prevents converging to bad local minima:
    - The trained small model (called “student”) is further constrained to mimic the functional properties of the teacher faithfully



✗: Cannot converge, as are not similar to $\hat{f}_{\text{large}}$

Agenda

- Which properties to be distilled?
- Which teacher to distill from?
- Other tips

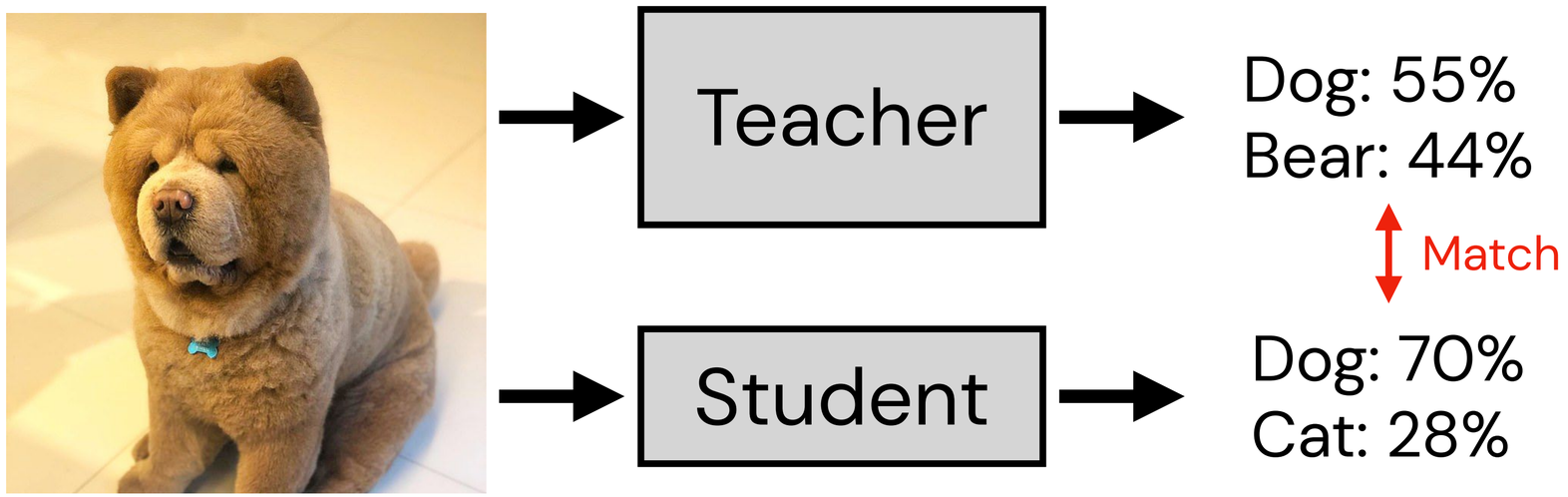
Which properties?

Which properties?

- Countless different forms
    - Logit
    - Feature
    - Relational

Logit distillation

- The classical form
- **Idea.** Make similar predictions at the logit level (a.k.a. “dark knowledge”)
    - Confidence, Close Classes, Wrong Predictions



Hinton et al., “Distilling the knowledge in a neural network,” arXiv 2015

Logit distillation

- Concretely, consider the K-class classification
  - **Logit.** Model output, before softmax (denoted by $z^{(1)}, \ldots, z^{(k)}$)
  - **(Generalized) Softmax.** The softmax is a probability

$$p^{(i)} = \frac{\exp(z^{(i)}/\tau)}{\sum_{j=1}^{k} \exp(z^{(j)}/\tau)}$$

    - τ: temperature hyperparameter
      - High temp → uniform (low confidence)
      - Low temp → zero/one (high confidence)

Hinton et al., "Distilling the knowledge in a neural network," arXiv 2015

Logit distillation

- The logit distillation minimizes the combined loss

$$\ell_{\mathrm{CE}} + \lambda \cdot \ell_{\mathrm{KD}}$$

  - ℓ_{CE}: Cross-entropy loss of the student

$$\ell_{\mathrm{CE}} = -\sum_{j=1}^{k} \mathbf{1}\{y = j\} \cdot \log p_{\mathrm{S}}^{(j)}$$

  - ℓ_{KD}: KL-divergence loss

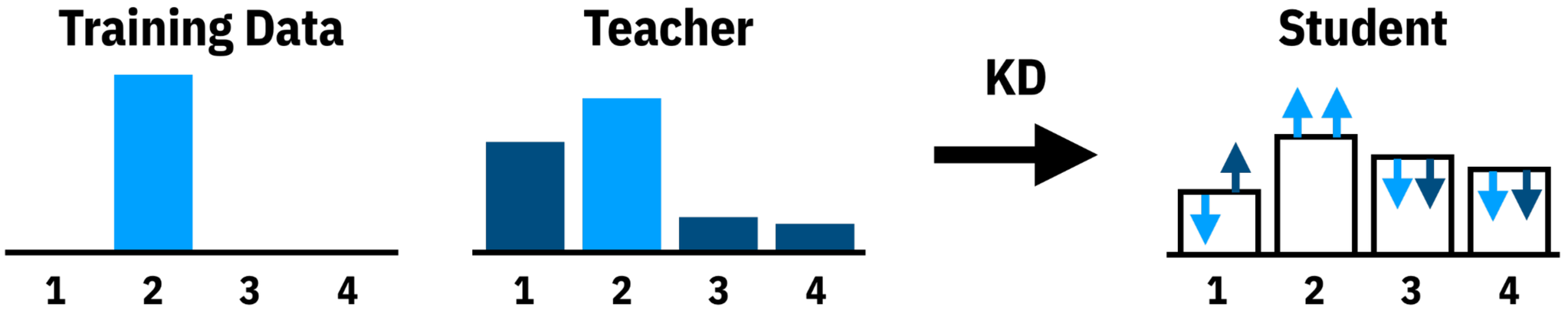
$$\ell_{\mathrm{KD}} = -\sum_{j=1}^{k} p_{\mathrm{T}}^{(j)} \log p_{\mathrm{S}}^{(j)}$$

    - $p_{\mathrm{T}}, p_{\mathrm{S}}$: teacher and student softmax, resp.

Hinton et al., "Distilling the knowledge in a neural network," arXiv 2015

Logit distillation

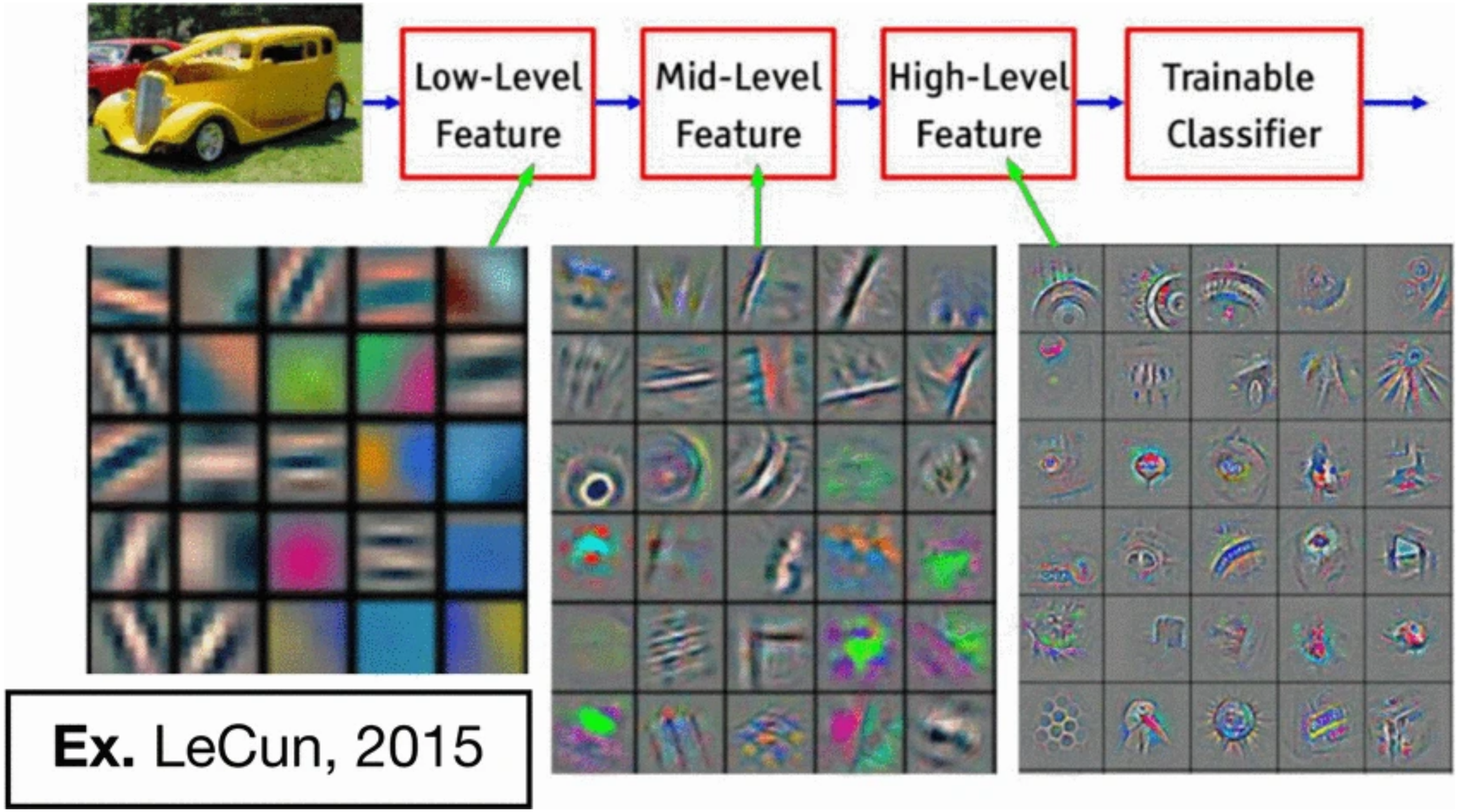
- **CE Loss.** Encourage student softmax to be close to zero-or-one
- **KD Loss.** Encourage student softmax to be similar to teacher softmax



Hinton et al., "Distilling the knowledge in a neural network," arXiv 2015

Feature distillation

- **Idea.** Regularize the teacher and student to have similar activations
  - That is, they make decisions based on similar reasons

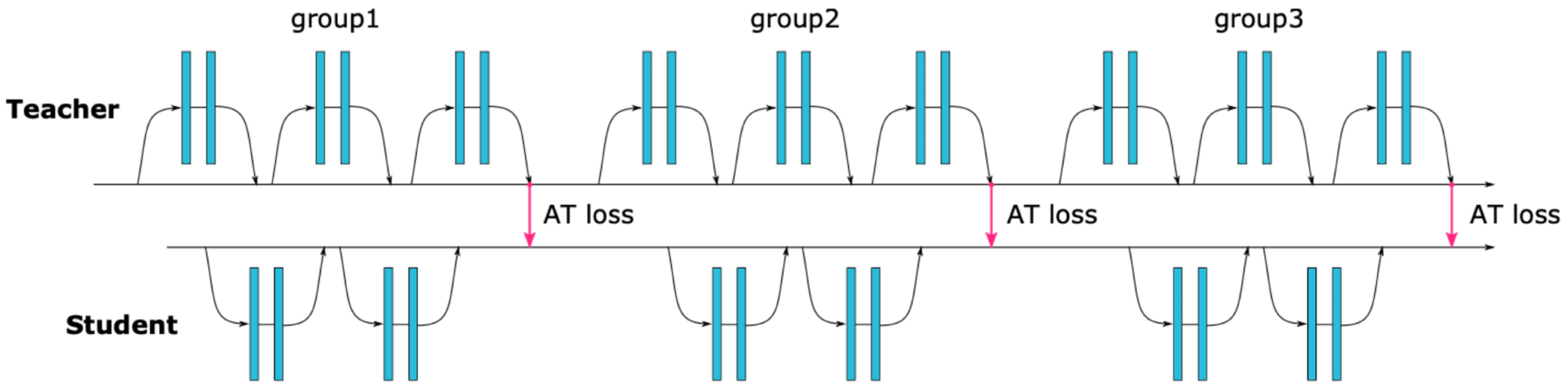


Feature distillation

- Feature distillation minimizes the combined loss

$$\ell_{\mathrm{CE}} + \lambda \cdot \sum_{j=1}^{J} \left\| \frac{Z_{\mathrm{T}}}{\|Z_{\mathrm{T}}\|} - \frac{Z_{\mathrm{S}}}{\|Z_{\mathrm{S}}\|} \right\|$$

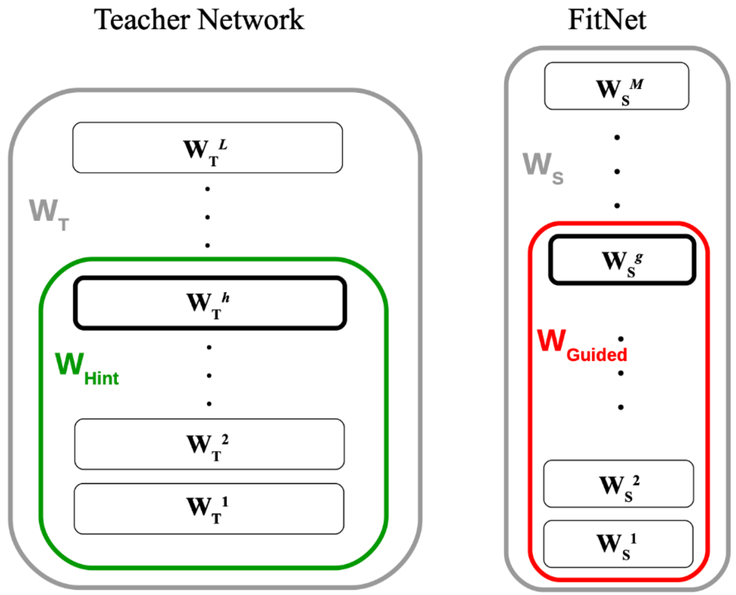
  - $Z_{\mathrm{T}}, Z_{\mathrm{S}}$: normalized activations of teacher and student (usually channel-pooled into HxW tensor)



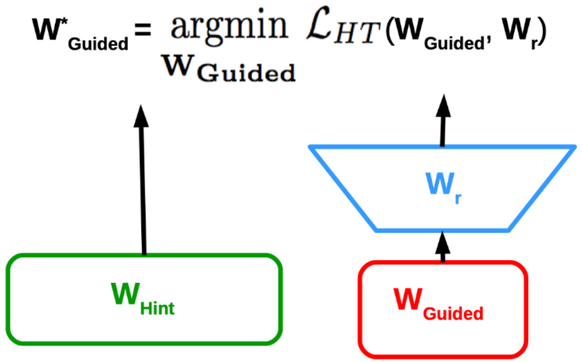
Zagoruyko and Komodakis, "Paying more attention to attention: Improving the performance of CNNs via Attention Transfer," ICLR 2017

Feature distillation

- Often, the activation sizes doesn't match
  - Small models typically downsample more aggressively
- **Solution.** Use additional projection (hint), and minimize the l2 difference



(a) Teacher and Student Networks

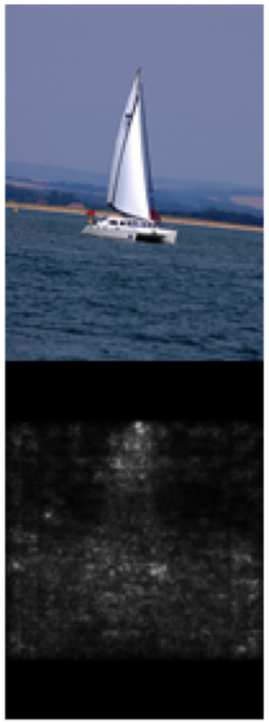


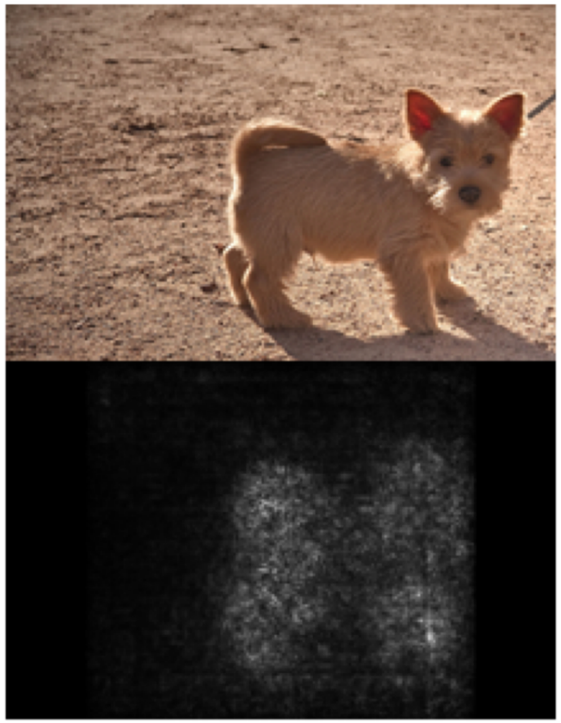
(b) Hints Training

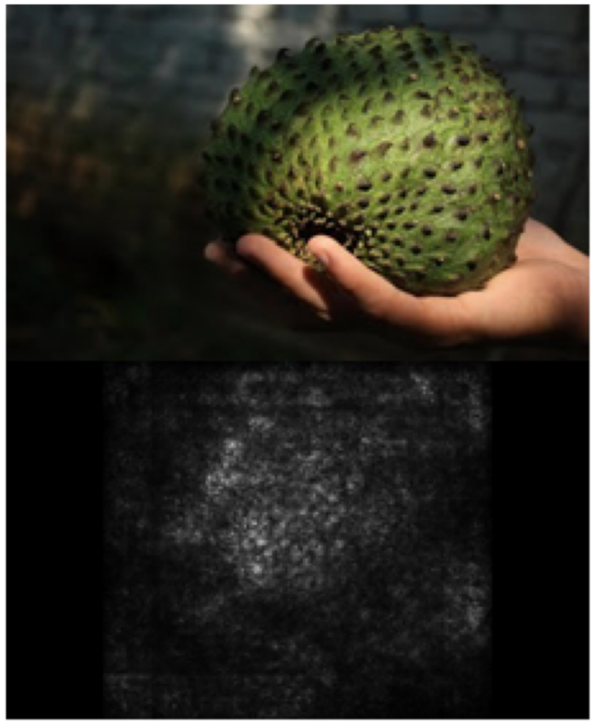
Romero et al., "FitNets: Hints for thin deep nets," ICLR 2015

Feature distillation

- Other variants match the input gradients (saliency map) of the models
  - Large gradient = The pixel affects the decision a lot
  - Models make predictions based on similar pixels



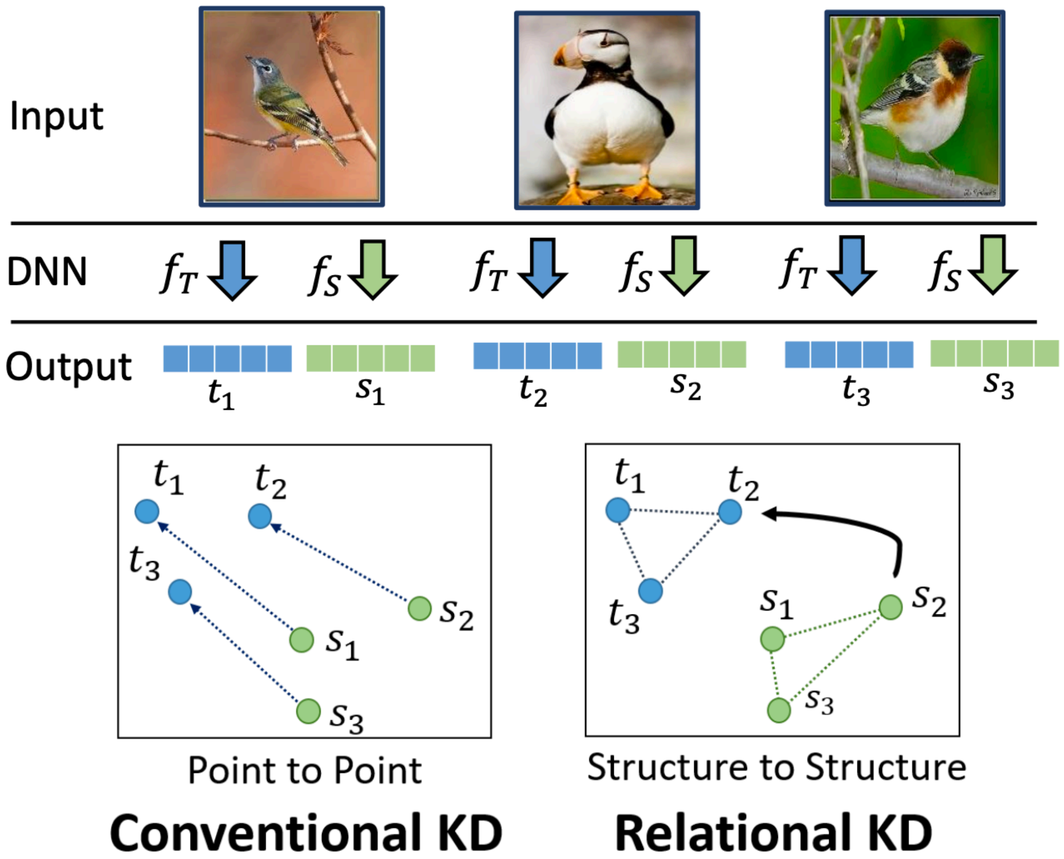




Simonyan et al., “Deep inside Convolutional Networks: Visualizing Image Classification Models and Saliency Maps,” ICLR workshop 2014

Relational distillation

- Some algorithms try to transfer inter-sample structures.
- **Idea.** If teacher thinks two samples are similar, the student should too
  - Angle between samples
  - Normalized distance



Park et al., "Relational Knowledge Distillation" CVPR 2019

Pros & cons

- **Logit**
    - Does not require internal information
        - thus applicable to closed-weight models
- **Feature, Relational**
    - Outperforms logit when carefully tuned
    - Natural choice for feature extractors

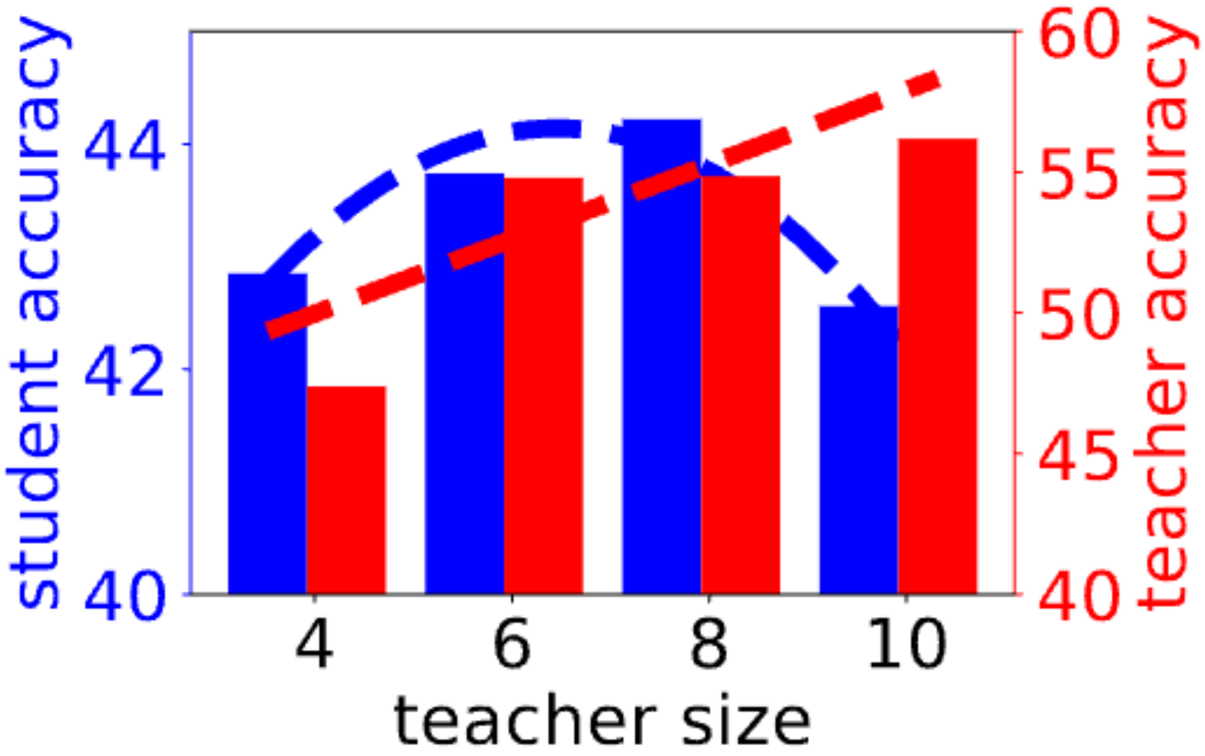
Which teacher?

Which teacher?

- Large model
    - Size discrepancy
- Ensemble of models
- Self-distillation
- Online distillation
- Uncompressed-to-compressed distillation

Size discrepancy

- **Observation.** If the teacher is too large, the KD does not work well
  - Intuition. The regularizer is too strong for the small model to fit



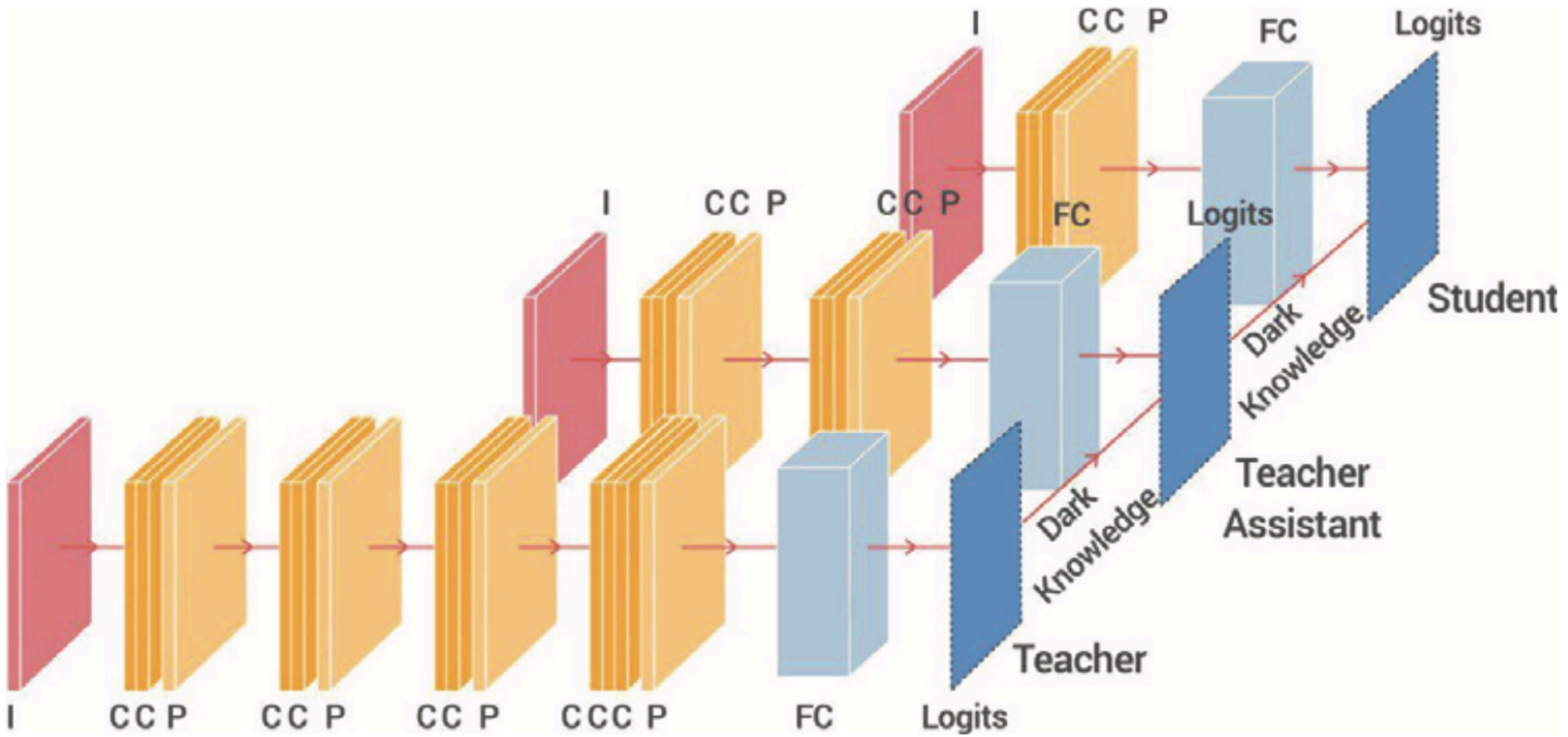
Mirzadeh et al., "Improved KD via Teacher Assistant" AAAI 2020

Size discrepancy

- **Solution.** Use the “Teacher Assistant”

  (1) Train an intermediate-sized TA, distilling teacher —> TA

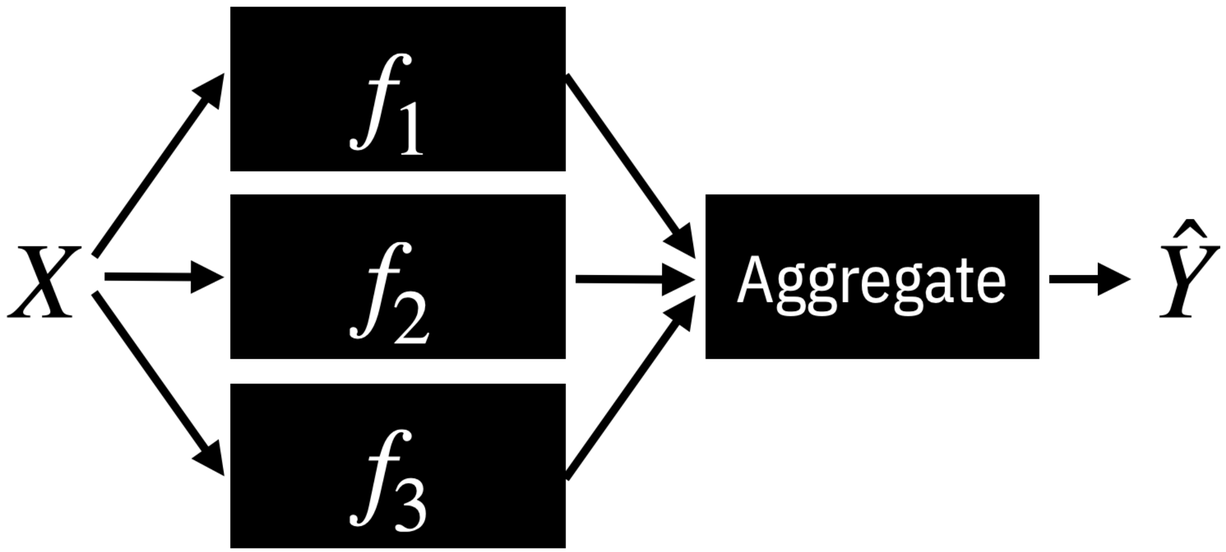
  (2) Train a student, distilling TA —> student



Mirzadeh et al., “Improved KD via Teacher Assistant” AAAI 2020

Ensemble of models

- The most classical way to attain a teacher is to form an ensemble
  - Make predictions on multiple weak models
    - Averaging
    - Majority voting



Ensemble of models

- **Brainteaser.** Suppose that we have 3 binary classifiers with 70% accuracy. What is the accuracy of the majority-voted-ensemble?

  (A) 70%

  (B) 78.4%

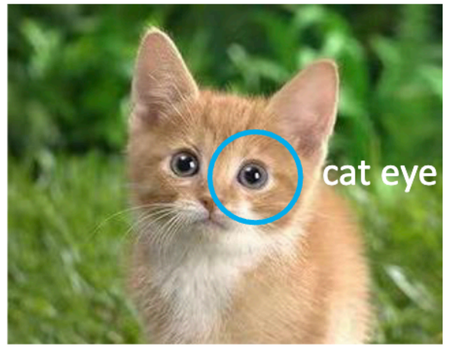
  (C) 100%

  (D) 55%

  (E) All of the above

Self-distillation

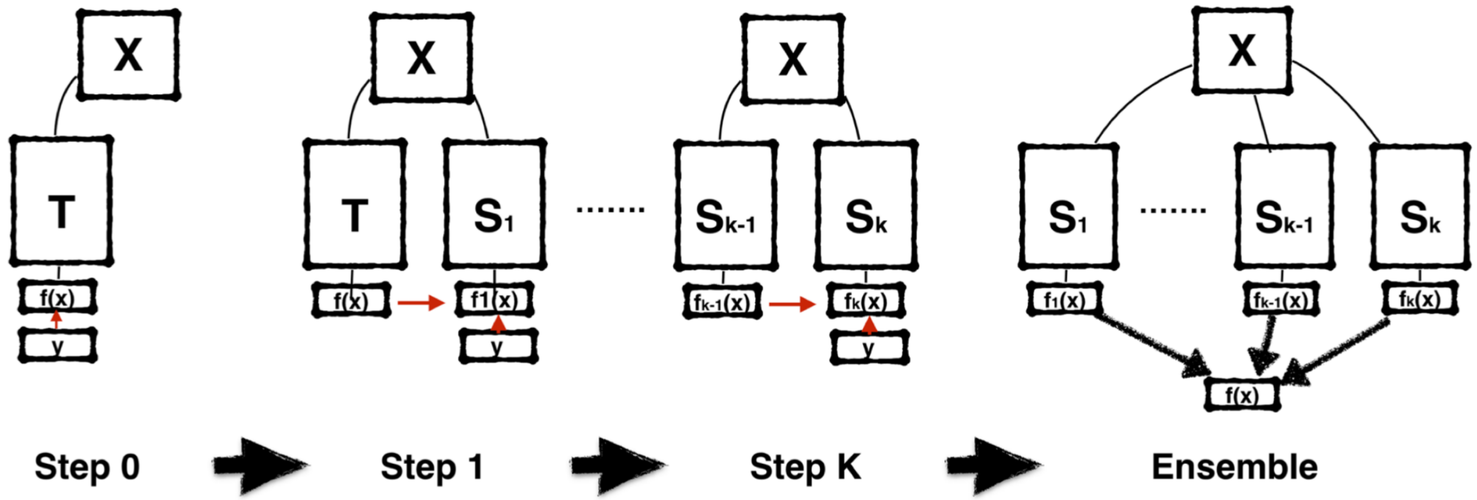
- **Fact.** The teacher of the same size as the student can improve the student
- **Why?** Different random seed captures different "views" of a data
  - With seed 1, the model learns to predict based on headlight
  - With seed 42, the model learns to predict based on window
    - After distillation, see headlight & window



Allen-Zhu and Li, "Towards Understanding Ensemble, Knowledge Distillation and Self-Distillation in Deep Learning," ICLR 2023

Self-distillation

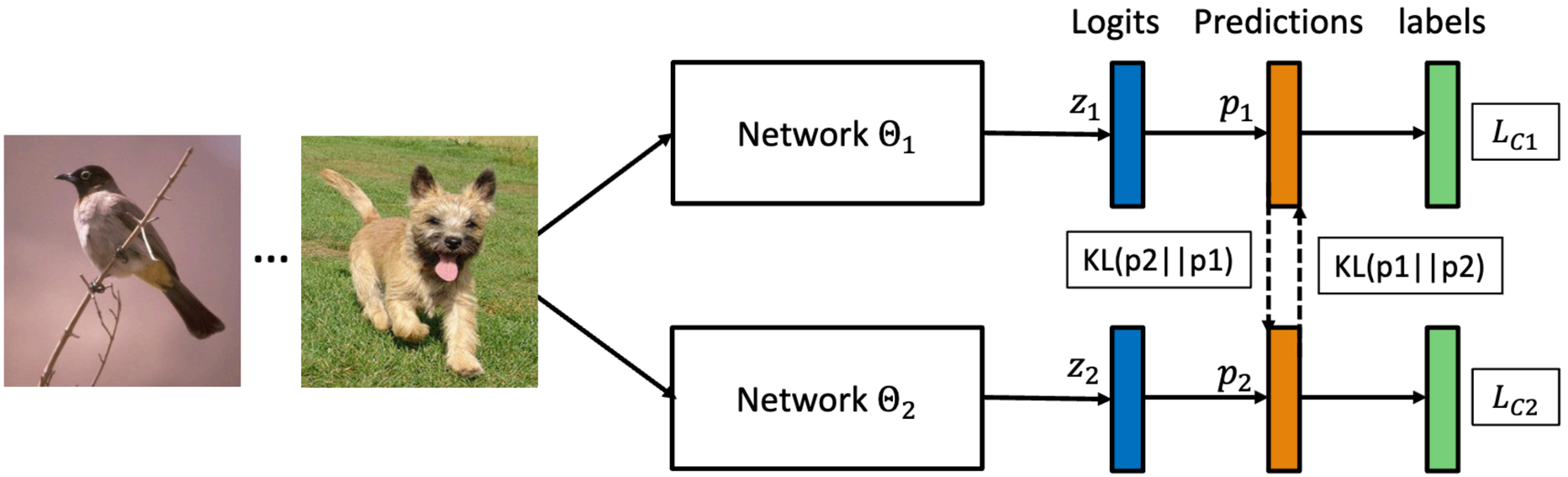
- **Born-again NN.** Distill from (an ensemble of) its prior versions
  - Train f_1 from scratch
  - Distill $f_1 \rightarrow f_2$
  - Distill $f_2 \rightarrow f_3$
  - (…)
  - Distill $f_{k-1} \rightarrow f_k$
  - Form an ensemble of $f_1, \ldots, f_k$



Furlanello et al., "Born again neural networks," ICML 2018

Online distillation

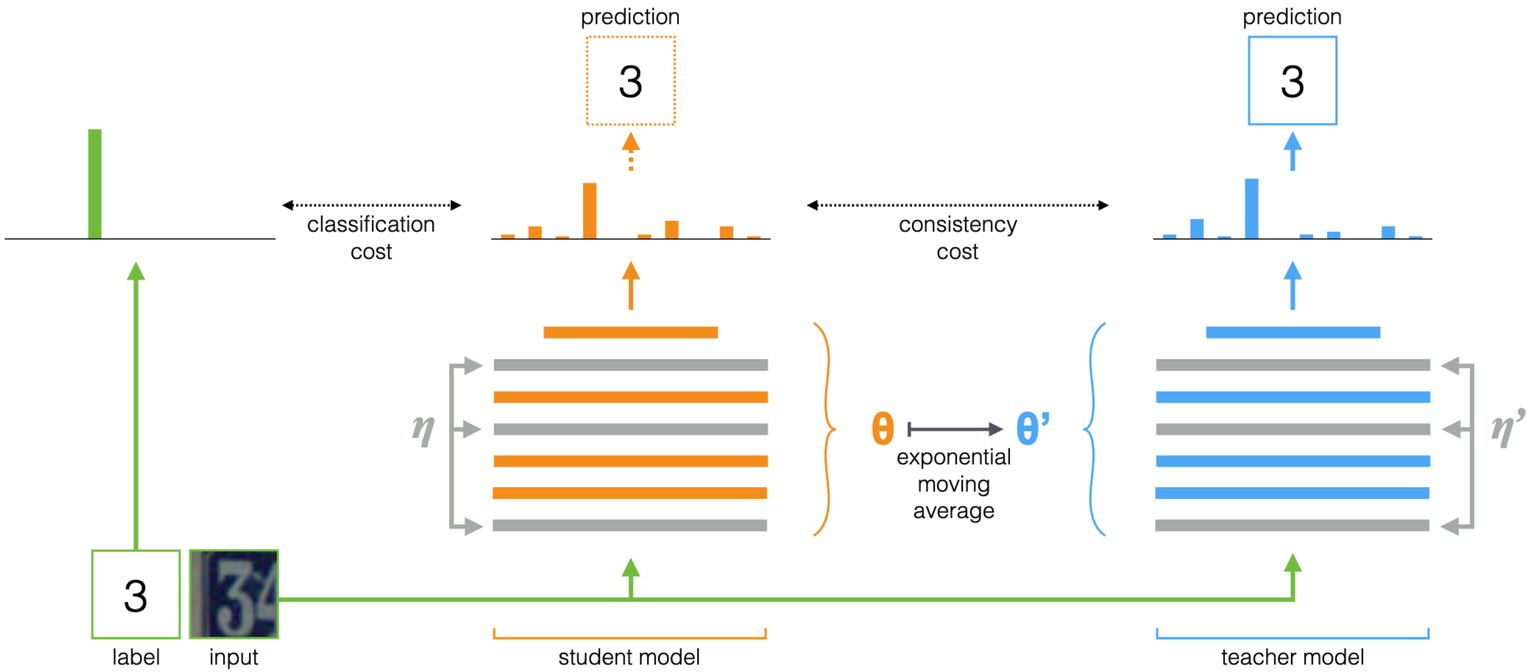
- **Idea.** One can train a teacher and a student simultaneously
  - Both networks distill from each other
  - No pretrained teacher required!



Zhang et al., "Deep mutual learning," CVPR 2018

Online distillation

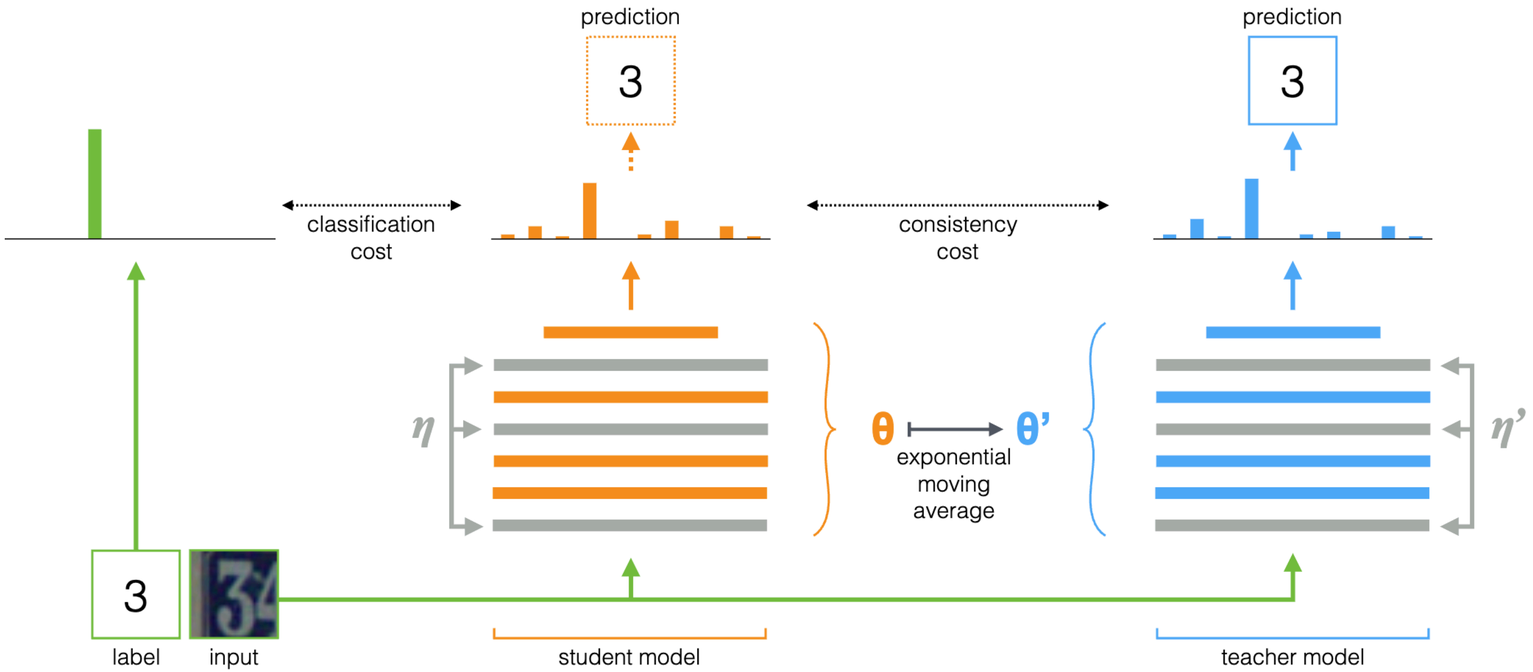
- Even further, we can use the momentum teacher
  - Use the exponential moving average of the student as the teacher (typically known to generalize better than original one)
  - Typically applied together with noise for self-supervised learning



Tarvainen and Valpola., "Mean teachers are better role models," NeurIPS 2017

Online distillation

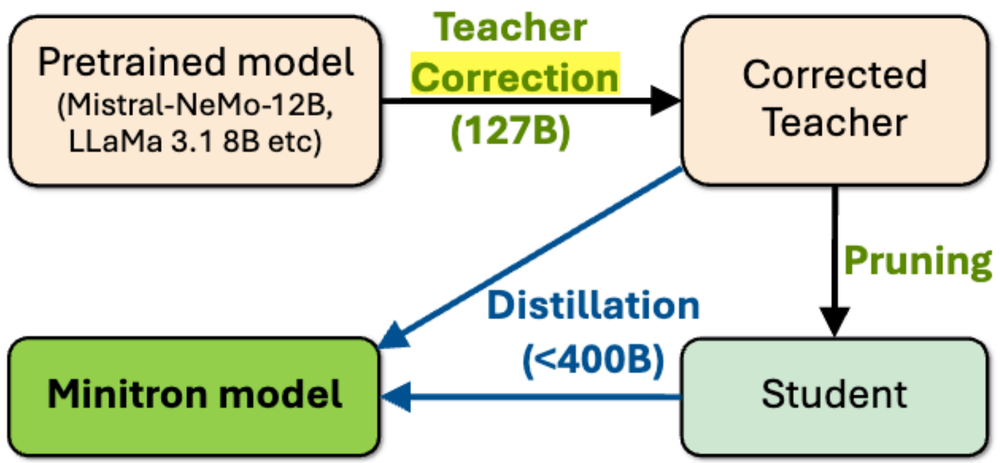
- Even further, we can use the momentum teacher
  - Use the exponential moving average of the student as the teacher (typically known to generalize better than original one)
  - Typically applied together with noise for self-supervised learning



Tarvainen and Valpola., "Mean teachers are better role models," NeurIPS 2017

Uncompressed-to-compressed

- A popular application is to combine with model compression
  - Example. Use dense model as teachers to retrain pruned models
    - **Teacher correction.** If original pretraining data is unavailable, fine-tuning on the distillation data is desirable (NVIDIA, 2024)



Sreenivas et al., "LLM Pruning and Distillation in Practice: The Minitron Approach," arXiv 2024

Other tips

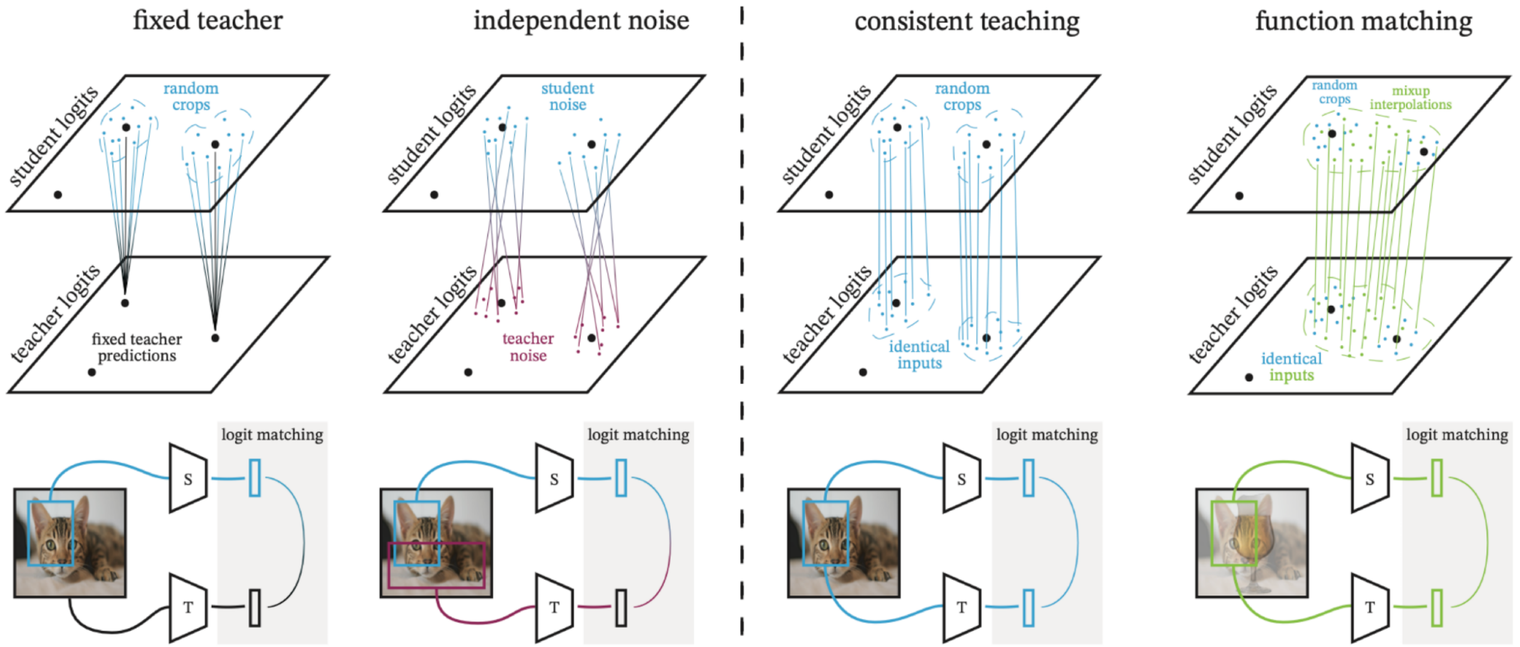
Other tips

- Two other aspects that affect the performance of the student
    - Data augmentation
    - Patience

Beyer et al., “KD: A Good Teacher is Patient and Consistent” ECCV 2022

Data augmentation

- It is essential that T & S look at the same views of the data
  - i.e., same augmented version



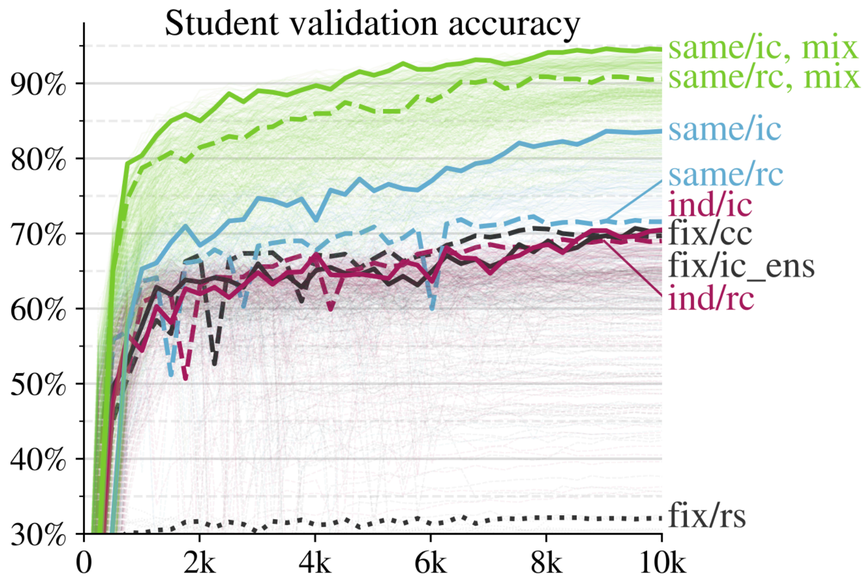
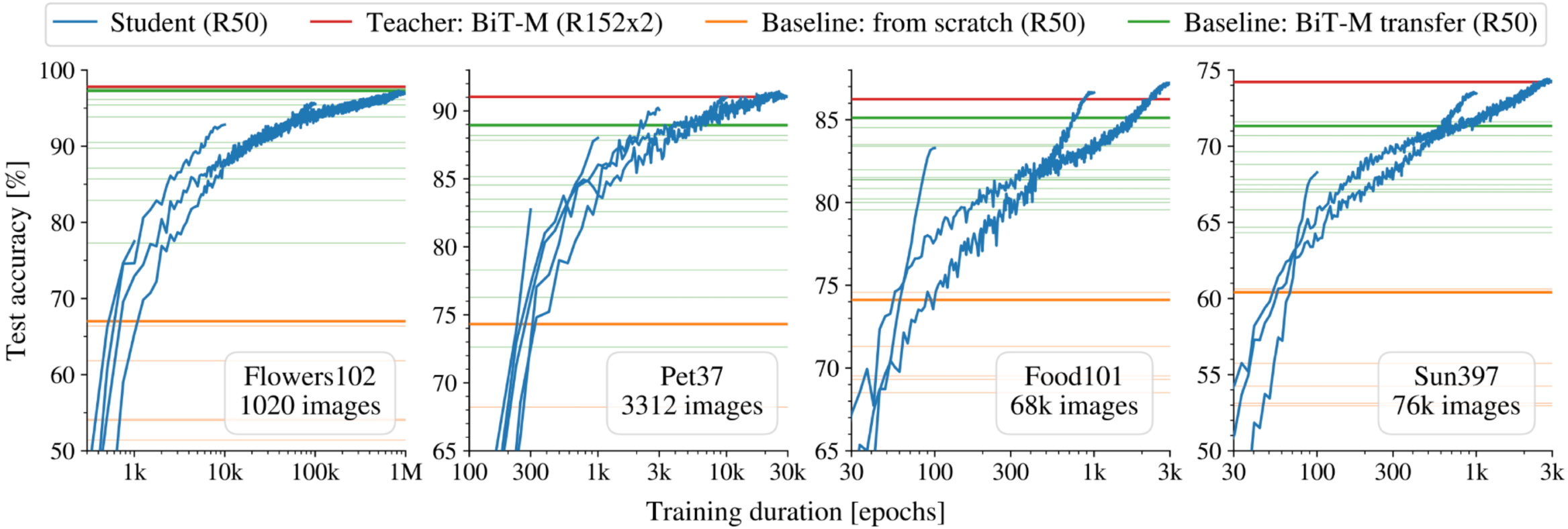


Figure 2. Schematic illustrations of various design choices when doing knowledge distillation. **Left**: *Teacher* receives a fixed image, while *student* receives a random augmentation. **Center-left**: *Teacher* and *student* receive independent image augmentations. **Center-right**: *Teacher* and *student* receive consistent image augmentations. **Right**: *Teacher* and *student* receive consistent image augmentations plus the input image manifold is extended by including linear segments between pairs of images (known as *mixup* [52] augmentation).

Beyer et al., "KD: A Good Teacher is Patient and Consistent" ECCV 2022

Patience

- It often takes very long time to attain the full benefit of distillation
  - c.f. grokking



Beyer et al., "KD: A Good Teacher is Patient and Consistent" ECCV 2022

Further readings

- **Distilling self-supervised models**
    - https://arxiv.org/abs/2101.04731
- **Data-free KD**
    - https://arxiv.org/abs/1904.01186

Next Class

- Neural Architecture Search

That’s it for today 🙌