Foundations of Edge AI

Lecture13

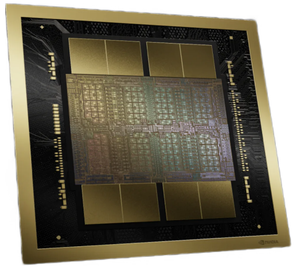
Knowledge Distillation

Lanyu (Lori) Xu

Email: lxu@oakland.edu
Homepage: https://lori930.github.io/
Office: EC 524

Limited Hardware Resources

NN must be tiny to run efficiently on tiny edge devices.



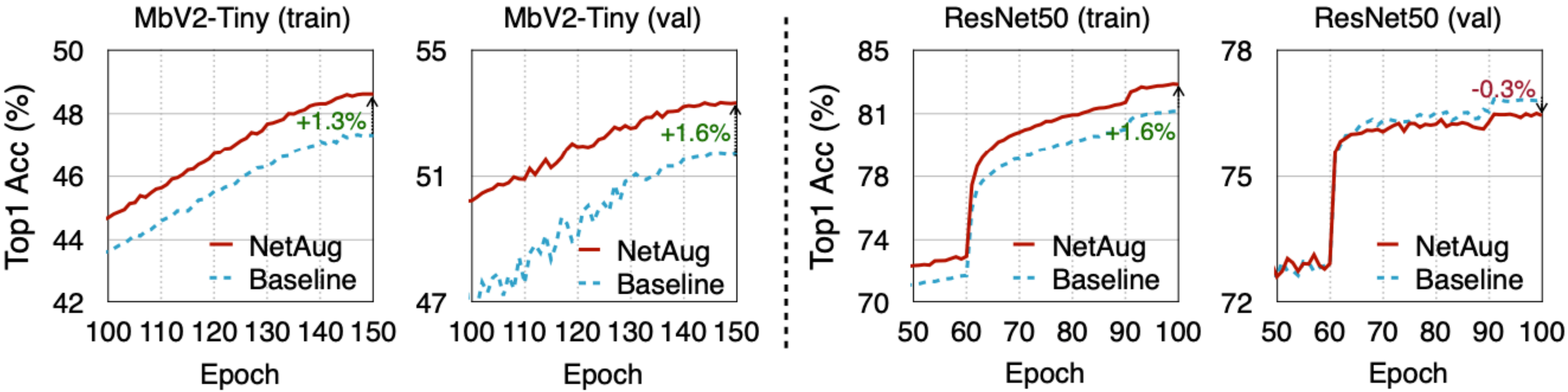
	Cloud AI	Mobile AI	Tiny AI
Memory (Activation)	80GB	~GB	320KB
Storage (Weights)	~TB/PB	~GB	1MB
Computation (FP32)	**10^3 TFLOPS**	**10-10^2 TFLOPS**	**MFLOPS**
Neural Network	**ResNet ViT-Large …**	**EfficientNet MobileNet …**	**MCUNet MobileNetV2-Tiny …**

Tiny Models are Hard to Train

Tiny models underfit large datasets

- Learning curves on ImageNet (Baseline)
  - Tiny model MobileNetV2 vs. “Large” model ResNet50

🤔 **Can we help the training of tiny models with large models?**

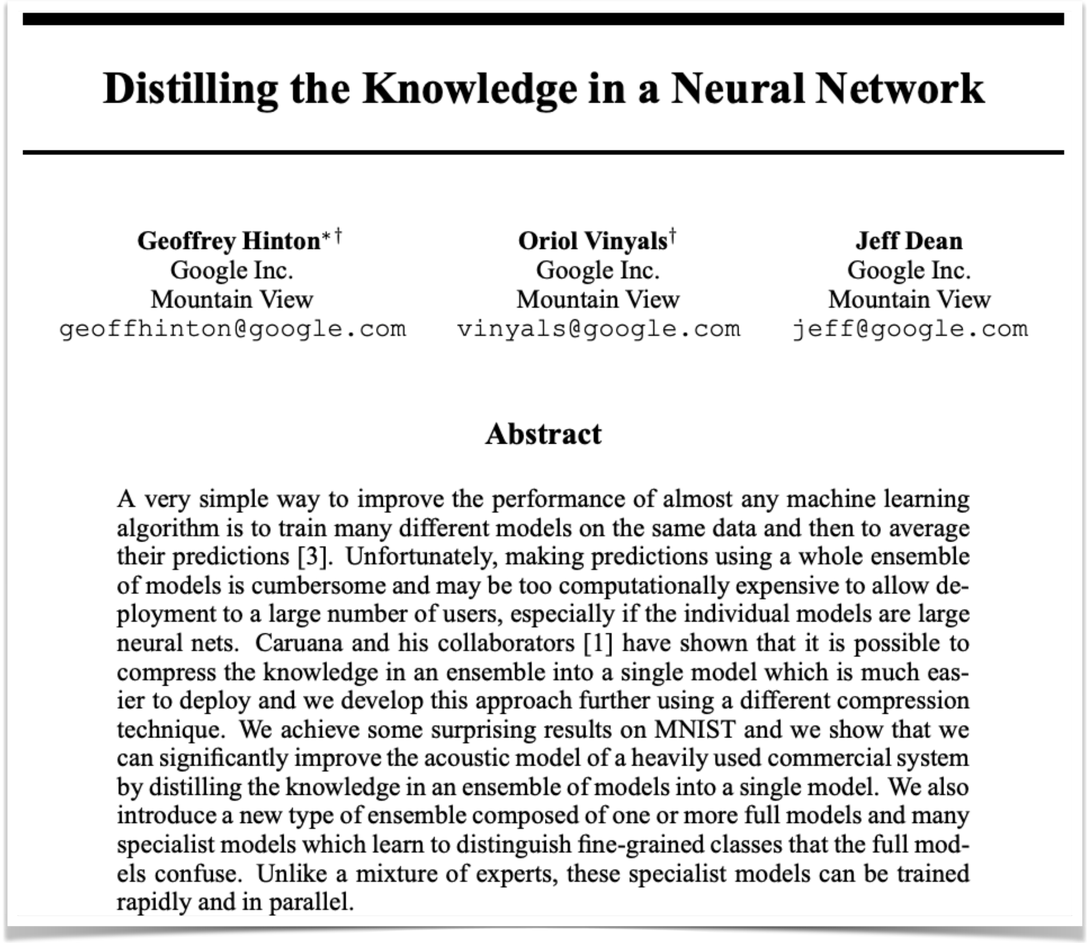


Cai, H., Gan, C., Lin, J., & Han, S. (2021). Network augmentation for tiny deep learning. arXiv preprint arXiv:2110.08890.

Lecture Plan

Today we will:

- Introduce knowledge distillation (KD)
  - What knowledge to distill
  - How to distill
  - Knowledge distillation applications
- An orthogonal way to KD: network augmentation for tiny DL



Distilling the Knowledge in a Neural Network

Geoffrey Hinton[*†]
Google Inc.
Mountain View
geoffhinton@google.com

Oriol Vinyals[†]
Google Inc.
Mountain View
vinyals@google.com

Jeff Dean
Google Inc.
Mountain View
jeff@google.com

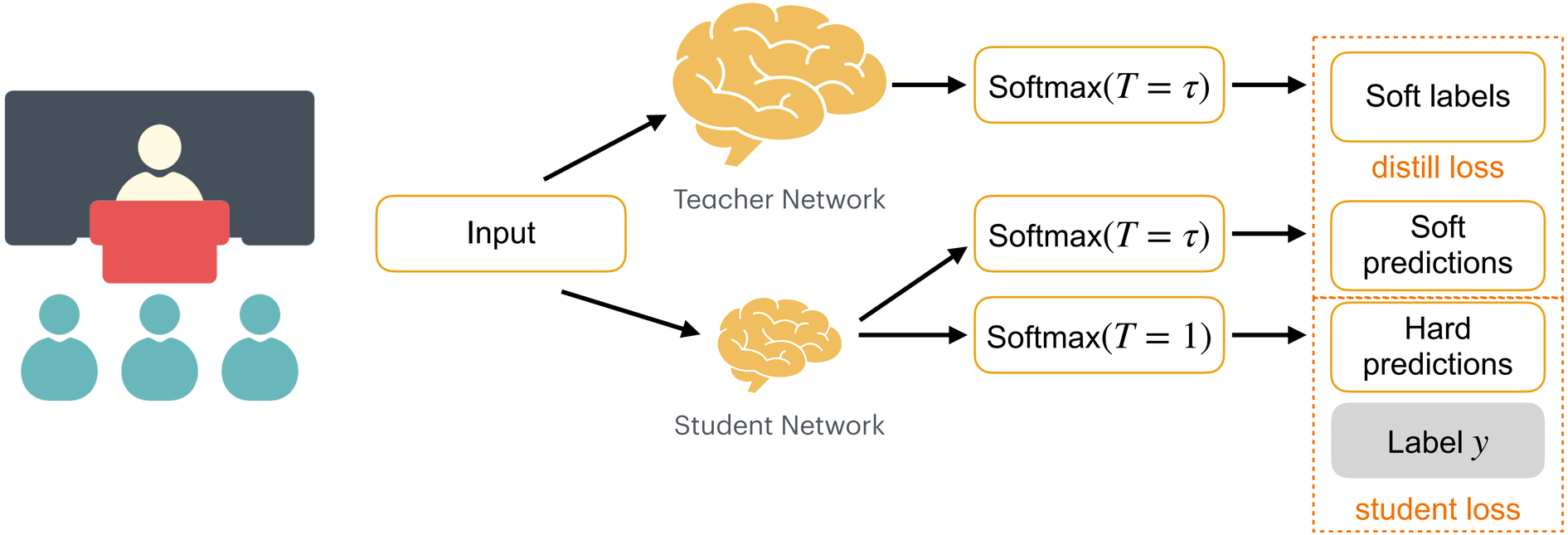
Abstract

A very simple way to improve the performance of almost any machine learning algorithm is to train many different models on the same data and then to average their predictions [3]. Unfortunately, making predictions using a whole ensemble of models is cumbersome and may be too computationally expensive to allow deployment to a large number of users, especially if the individual models are large neural nets. Caruana and his collaborators [1] have shown that it is possible to compress the knowledge in an ensemble into a single model which is much easier to deploy and we develop this approach further using a different compression technique. We achieve some surprising results on MNIST and we show that we can significantly improve the acoustic model of a heavily used commercial system by distilling the knowledge in an ensemble of models into a single model. We also introduce a new type of ensemble composed of one or more full models and many specialist models which learn to distinguish fine-grained classes that the full models confuse. Unlike a mixture of experts, these specialist models can be trained rapidly and in parallel.

Hinton, G. (2015). Distilling the Knowledge in a Neural Network. arXiv preprint arXiv:1503.02531.

Illustration of Knowledge Distillation

Matching prediction probabilities between teacher and student



- Neural networks typically produce class probabilities by using a "softmax" output layer that converts the logit, z_i, computed for each class into a probability, q_i, by comparing z_i with the other logits:

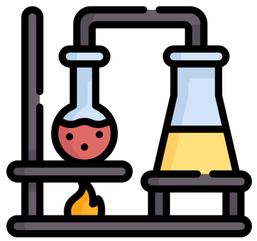
$$q_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)}$$

Hinton, G. (2015). Distilling the Knowledge in a Neural Network. arXiv preprint arXiv:1503.02531.

Illustration of Knowledge Distillation

Matching prediction probabilities between teacher and student

$$q_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)}$$



Teacher Network

Student Network

	Logits	Probabilities
Cat	5	0.982
Dog	1	0.017

$$= \frac{\exp(5)}{\exp(5) + \exp(1)}$$

$$= \frac{\exp(1)}{\exp(5) + \exp(1)}$$

	Logits	Probabilities
Cat	3	0.731
Dog	2	0.269

The teacher model is with very high confidence;

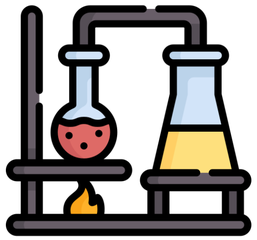
vs.

The student model is less confident.

Illustration of Knowledge Distillation

Matching prediction probabilities between teacher and student

- Introduce the concept of **temperature** T



$$q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

Teacher Network

	Logits	Probabilities (T=1)	Probabilities (T=10)
Cat	5	0.982	0.599
Dog	1	0.017	0.401

$$= \frac{\exp(5)}{\exp(5) + \exp(1)} \qquad = \frac{\exp(5/10)}{\exp(5/10) + \exp(1/10)}$$

A more confident and peaked around the predicted class

Softer output probabilities to provide richer information

- A higher value for T smooths the output probability distribution.

Hinton, G. (2015). Distilling the Knowledge in a Neural Network. arXiv preprint arXiv:1503.02531.

What is Knowledge Distillation

- **Intuition**: To train a model that can generalize well requires information about the correct way to generalize. When distilling the knowledge from a large model into a small one, we can train the small model to generalize in the same way as the large model.

- To transfer the generalization ability, one obvious way is to use the probability produced by the large model as “soft targets” for training the small model.

- One general solution to this is **distillation**: raise the temperature of the final softmax until the large model produces a suitably soft set of targets, then use the same high temperature when training the small model to match these soft targets.

- When computing the logits, we scale the softmax function with the distillation temperature T to **align the class probability distributions between teacher and student networks.**

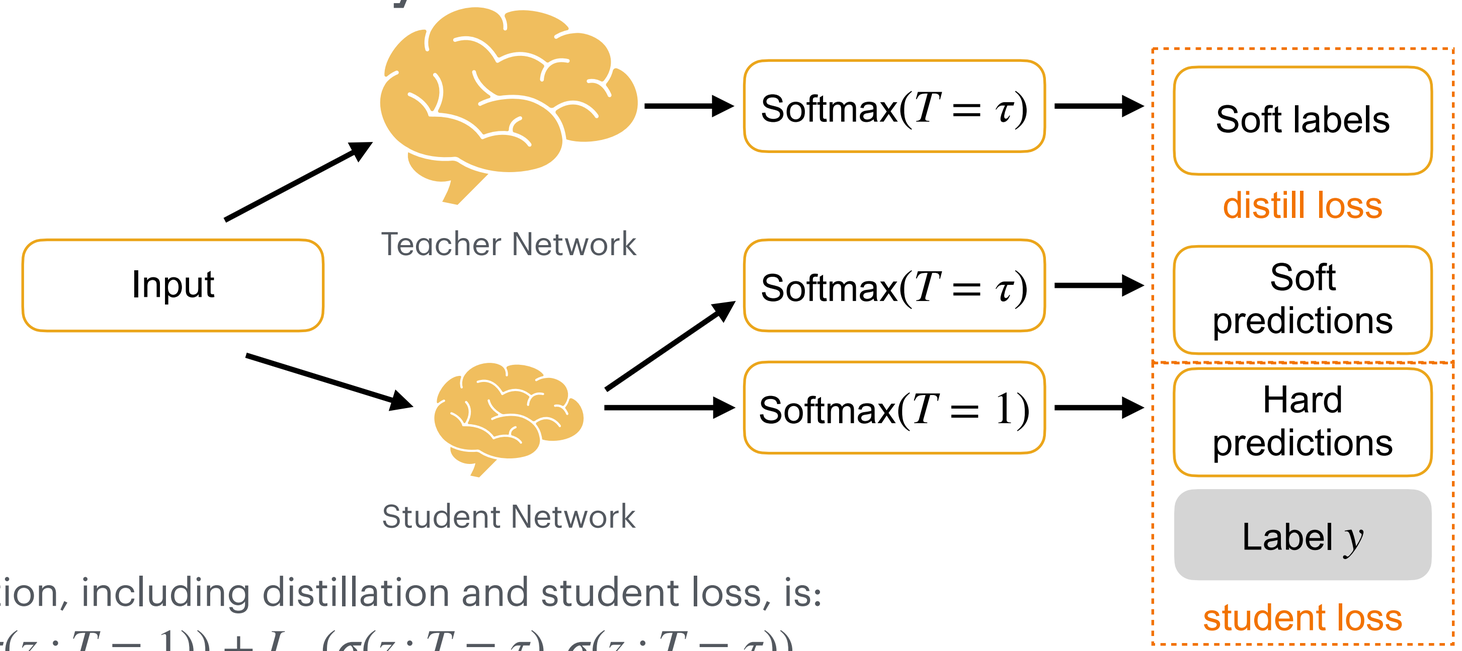
$$q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

Hinton, G. (2015). Distilling the Knowledge in a Neural Network. arXiv preprint arXiv:1503.02531.

What Knowledge to Distill?

Distill Output Logits

The simplest and widely used



Cross entropy loss

$$\mathbb{E}(-p_t \log p_s)$$

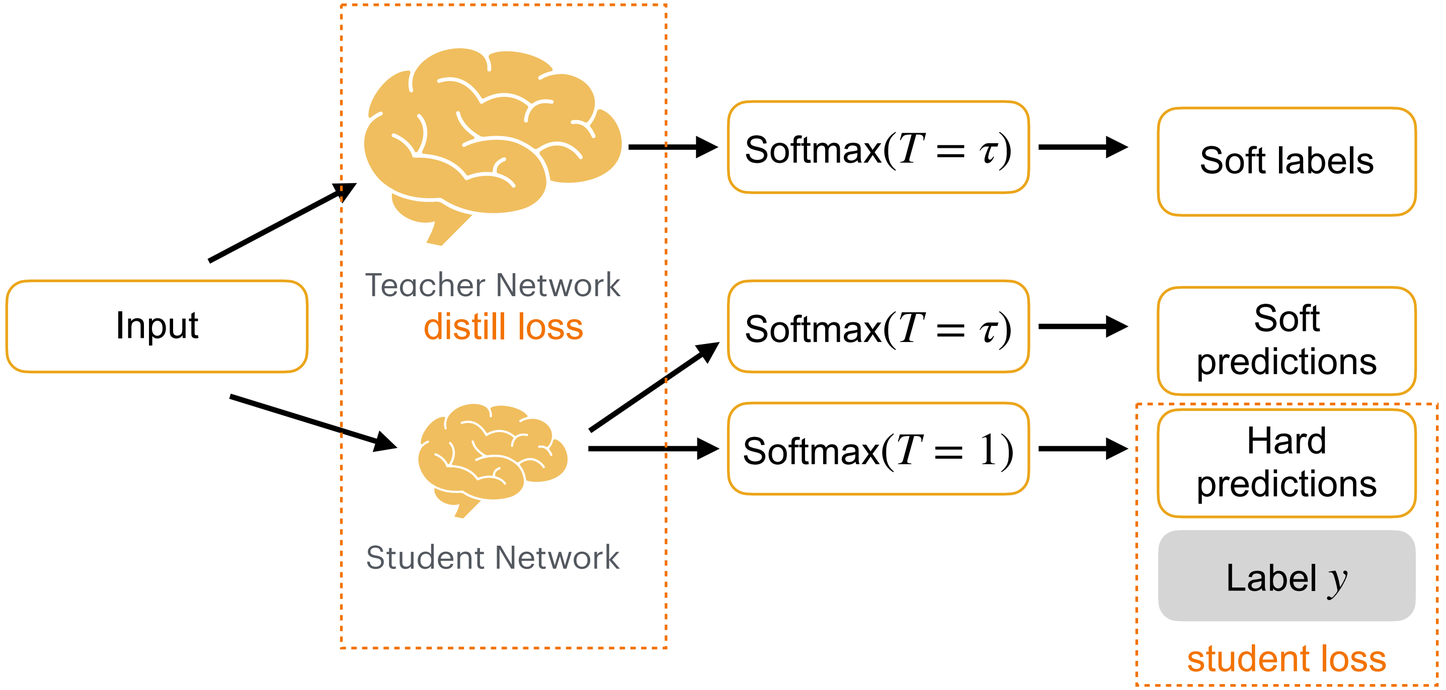
L2 loss

$$\mathbb{E}(\| p_t - p_s \|_2^2)$$

- The overall loss function, including distillation and student loss, is:
  $L(x;\theta) = \alpha * L_{ce}(y, \sigma(z_s; T = 1)) + L_{ce}(\sigma(z_t; T = \tau), \sigma(z_s; T = \tau))$
  - where x is the input, θ is the student network parameter, L_{ce} is the cross-entropy loss function, y is the ground truth label, σ is the softmax function parameterized by the temperature T (controls probability distribution). α is the coefficient. z_s, z_t are the logits of the students and teacher respectively.
  - In other words, the loss function is a weighted average of cross-entropy loss and KL divergence.

Hinton, G. (2015). Distilling the Knowledge in a Neural Network. arXiv preprint arXiv:1503.02531.
Ba, J., & Caruana, R. (2014). Do deep nets really need to be deep?. Advances in neural information processing systems, 27.

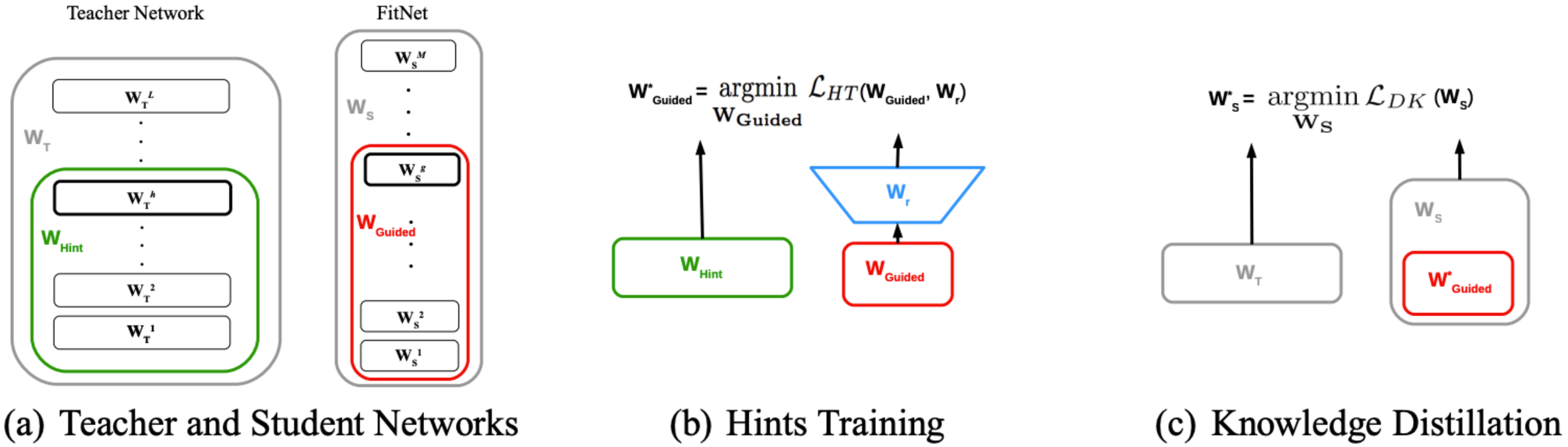
Distill Intermediate Weights



- Can we directly calculate the L2 norm distance between the teacher's weight and the student's weight?

Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2021). Knowledge distillation: A survey. International Journal of Computer Vision, 129(6), 1789-1819.

Distill Intermediate Weights



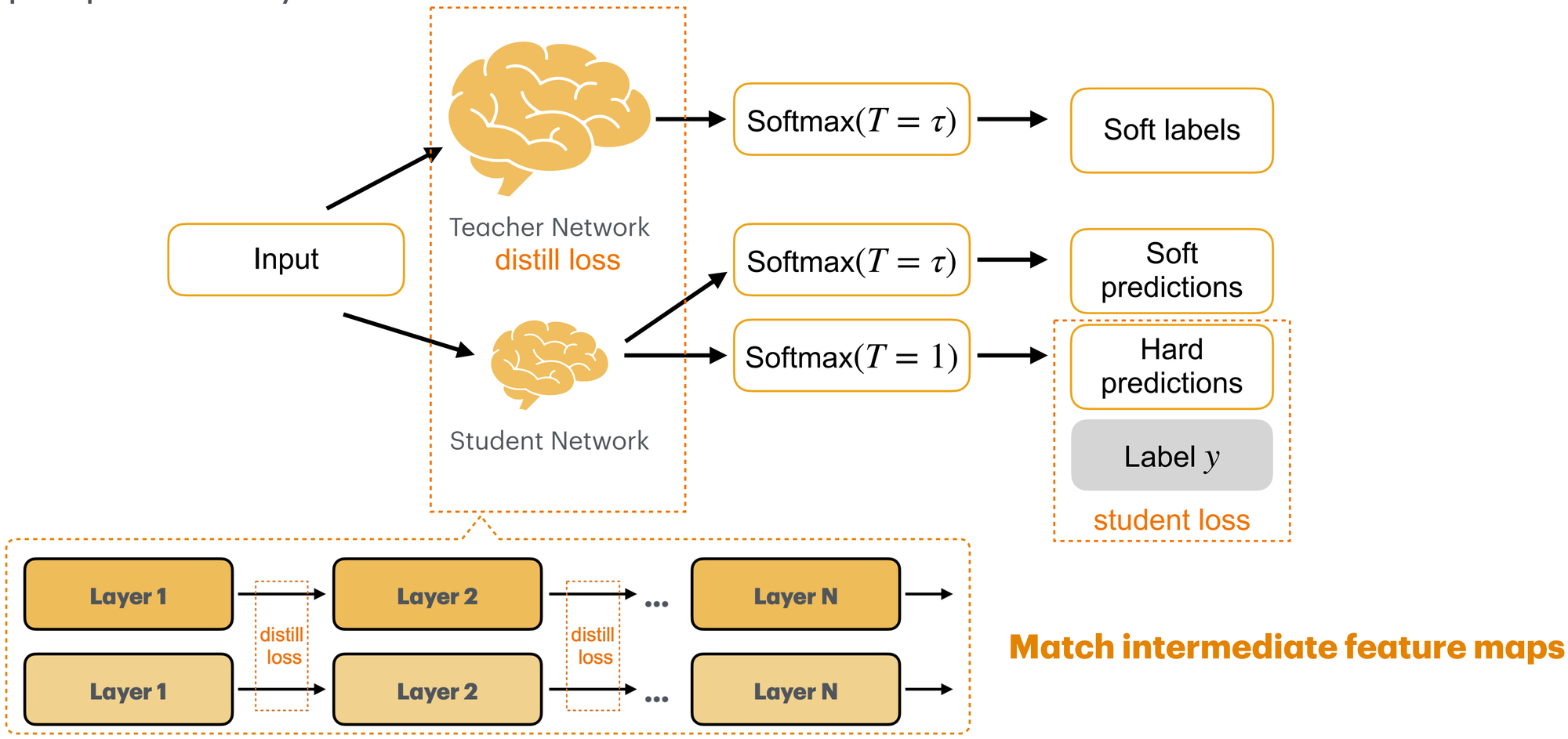
(a) Teacher and Student Networks (b) Hints Training (c) Knowledge Distillation

- Other than the cross-entropy distillation loss $\mathscr{L}_{KD}(W_s)$, also add an L2 loss between teacher weights and student weights $\mathscr{L}_{HT}(W_{Guided}, W_r)$
  - Linear transformation (W_r) is applied to match the dimensionalities (blue block)
- If the teacher's channel is 512, student's channel is 256, what is the dimension of W_r?

Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., & Bengio, Y. (2014). Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550.

Distill Intermediate Features

- **Intuition**: teacher and student networks should have similar **feature** distributions, not just output probability distributions

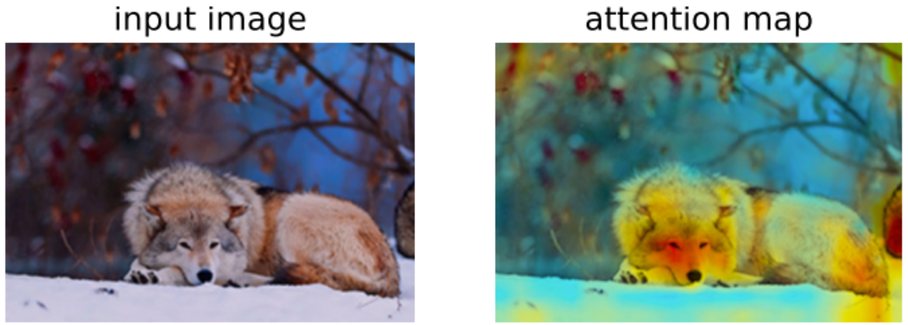


Huang, Z., & Wang, N. (2017). Like what you like: Knowledge distill via neuron selectivity transfer. arXiv preprint arXiv:1707.01219.

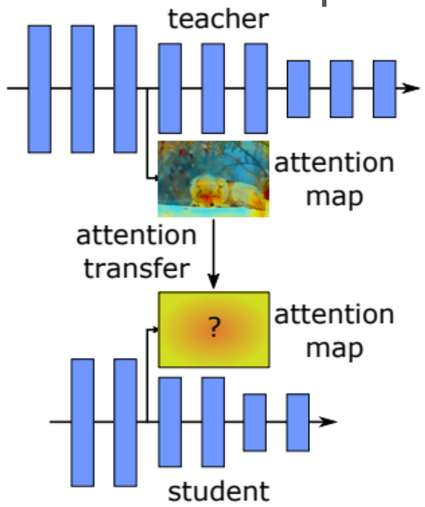
Distill Intermediate Attention Maps

Gradients of feature maps are used to characterize "attention" of DNNs

- The attention of a CNN feature map x is defined as $\frac{\partial L}{\partial x}$, where L is the learning objective
- **Intuition**: if $\frac{\partial L}{\partial x_{i,j}}$ is large, a small pertubation at i, j will significantly impact the final output. As the result, the network is putting more attention on position i, j



Input image and its attention map



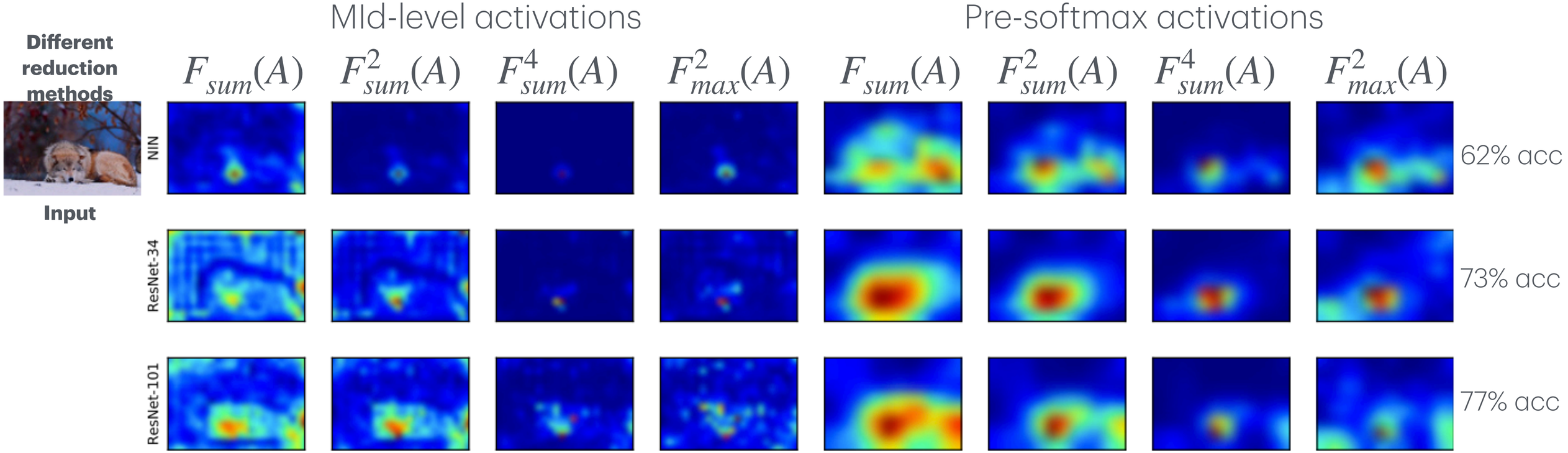
Attention transfer

Zagoruyko, S., & Komodakis, N. (2016). Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928.

Distill Intermediate Attention Maps

Performant models have similar attention maps

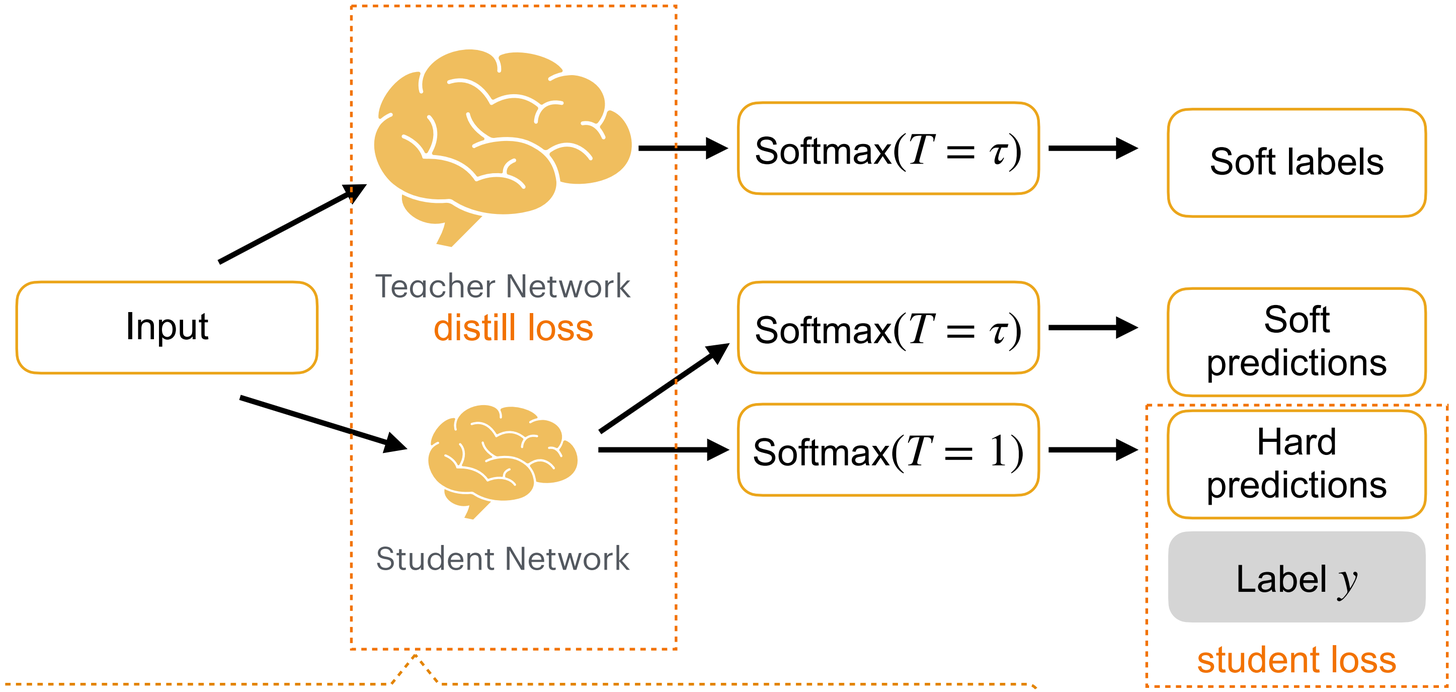
- Attention maps of performant ImageNet models (ResNets) are indeed similar to each other, but the less performant model (NIN) has quite different attention maps

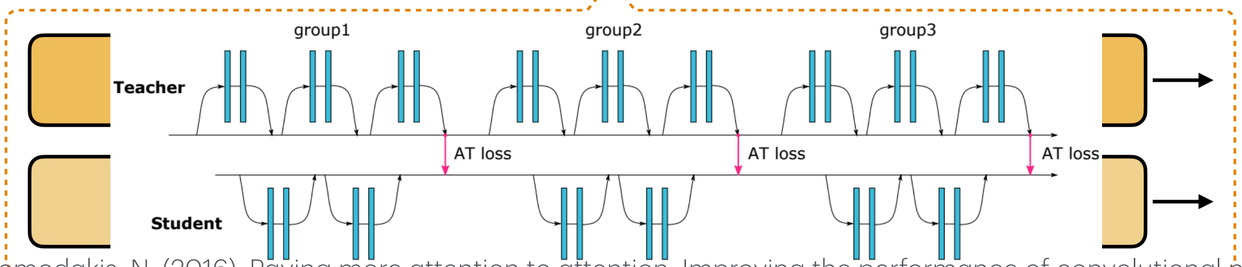


Zagoruyko, S., & Komodakis, N. (2016). Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928.

Distill Intermediate Attention Maps

- Intuition: teacher and student networks should have similar **feature** distributions, not just output probability distributions



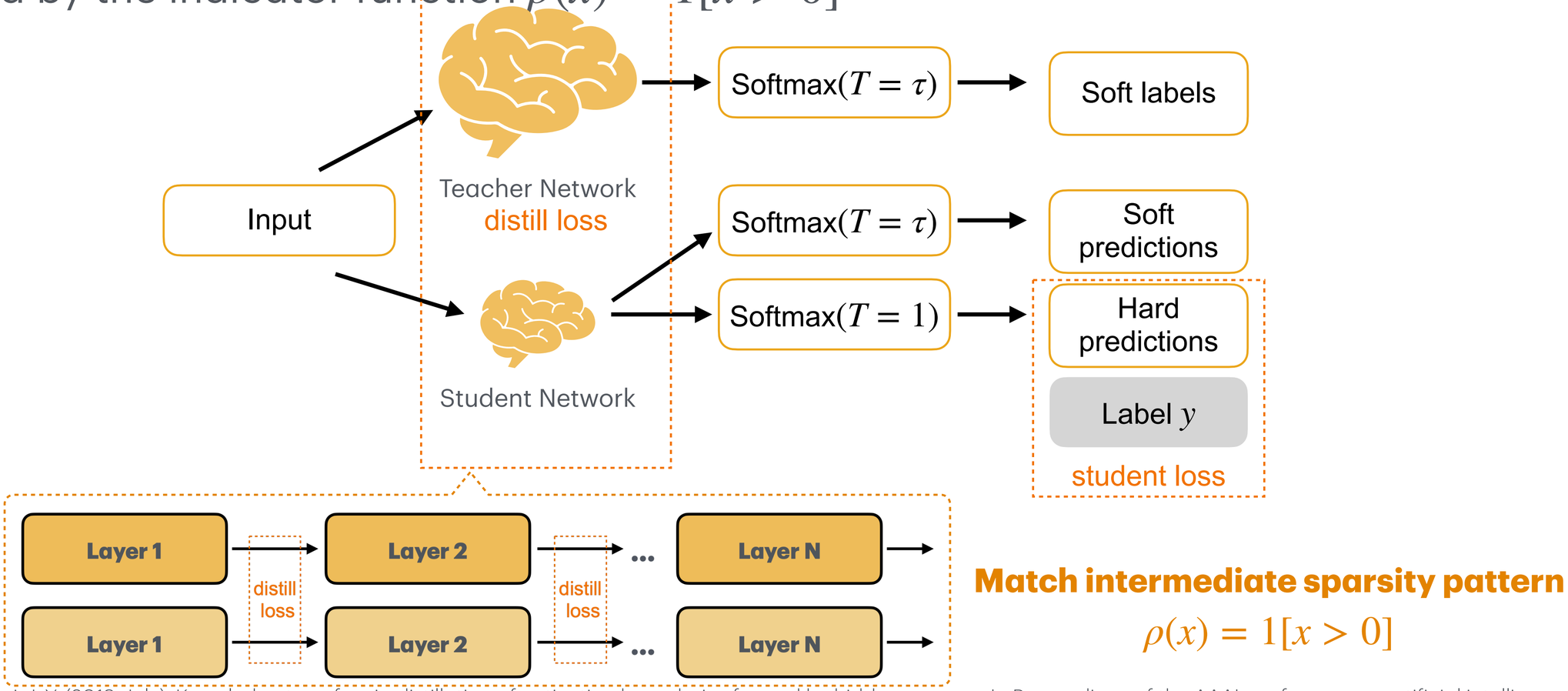


Match intermediate attention maps $\frac{\partial L}{\partial x}$

Zagoruyko, S., & Komodakis, N. (2016). Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928.

Distill Sparsity Patterns

- **Intuition**: the teacher and student networks should have similar sparsity patterns after the **ReLU activation.** A neuron is activated after ReLU if its value is larger than 0, denoted by the indicator function $\rho(x) = 1[x > 0]$



Match intermediate sparsity pattern

$$\rho(x) = 1[x > 0]$$

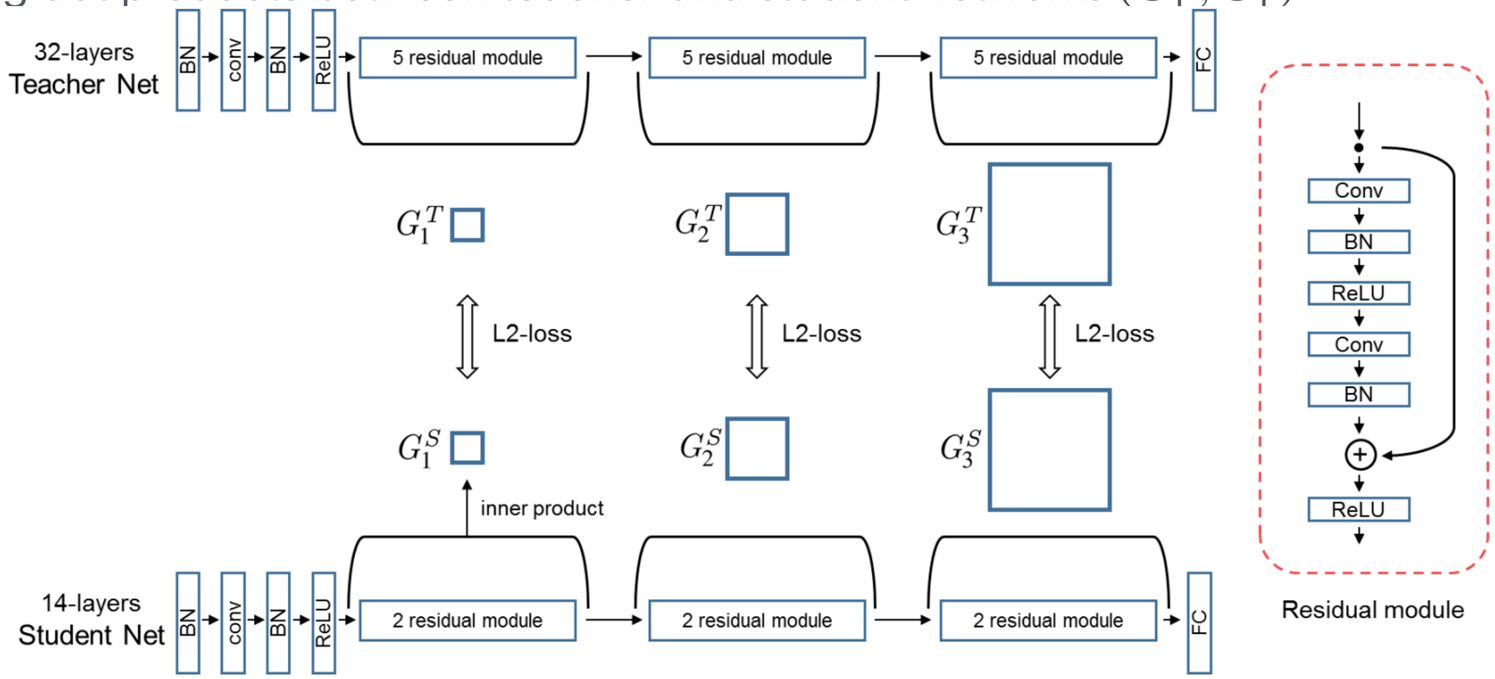
Heo, B., Lee, M., Yun, S., & Choi, J. Y. (2019, July). Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 3779-3787).

Distill Relational Information

Relations between different layers

- Use inner produce to extract relational information (a matrix of shape $C_i \times C_o$, reduction on the spatial dimensions) for both student and teacher networks

- Note: the student and teacher networks only **differ in the number of layers, not the number of channels**. Then match the resulting dot products between teacher and student networks (G_1^T, G_1^S)

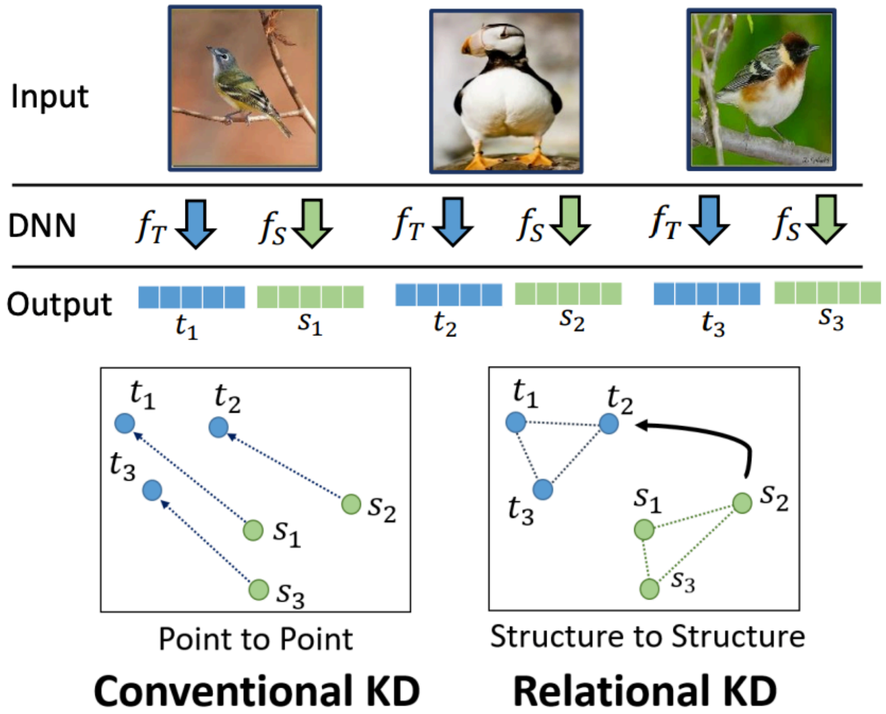


Yim, J., Joo, D., Bae, J., & Kim, J. (2017). A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4133-4141).

Distill Relational Information

Relations between different samples

- Conventional KD focuses on matching features/logits for **one input**.
- Relational KD looks at the **relations between intermediate features** from **multiple inputs**.

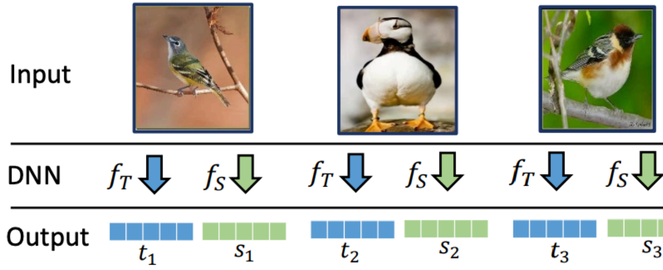


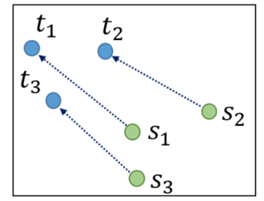
Park, W., Kim, D., Lu, Y., & Cho, M. (2019). Relational knowledge distillation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3967-3976).

Distill Relational Information

Relations between different samples

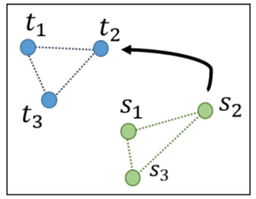
- Conventional KD focuses on matching features/logits for **one input**.
- Relational DK looks at the **relations between intermediate features** from **multiple inputs**.
- $\psi(s_1, s_2, \ldots, s_n) = (\| s_1 - s_2 \|_2^2, \| s_1 - s_3 \|_2^2, \ldots, \| s_1 - s_n \|_2^2, \ldots \| s_{n-1} - s_n \|_2^2)$ is a vector of length $\frac{n(n-1)}{2}$ representing pairwise distances of feature vectors





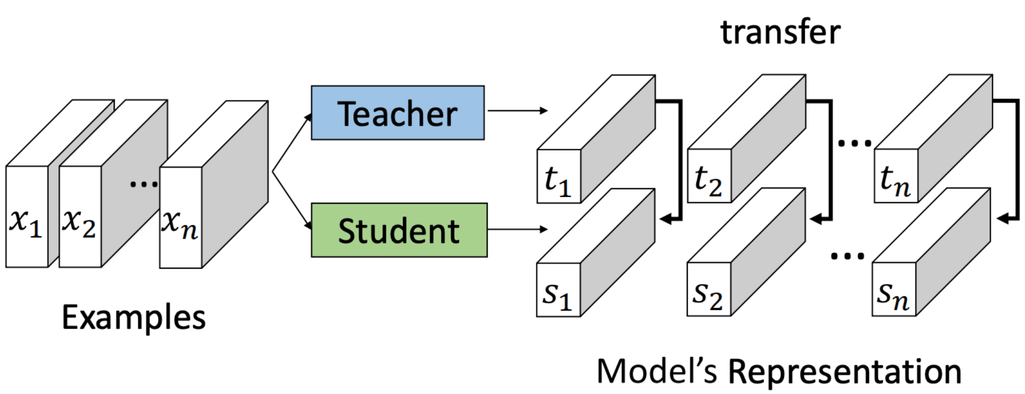
Point to Point

Conventional KD

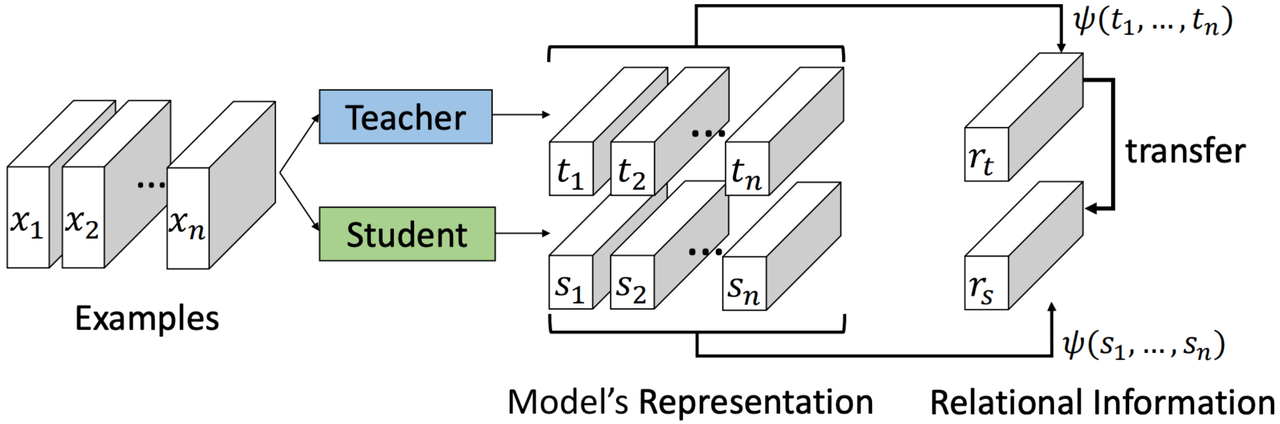


Structure to Structure

Relational KD



Individual Knowledge Distillation



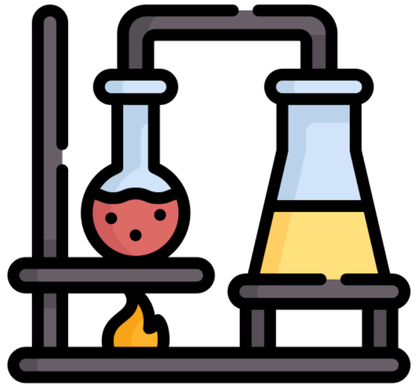
Relational Knowledge Distillation

Park, W., Kim, D., Lu, Y., & Cho, M. (2019). Relational knowledge distillation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3967-3976).

What Knowledge to Distill?

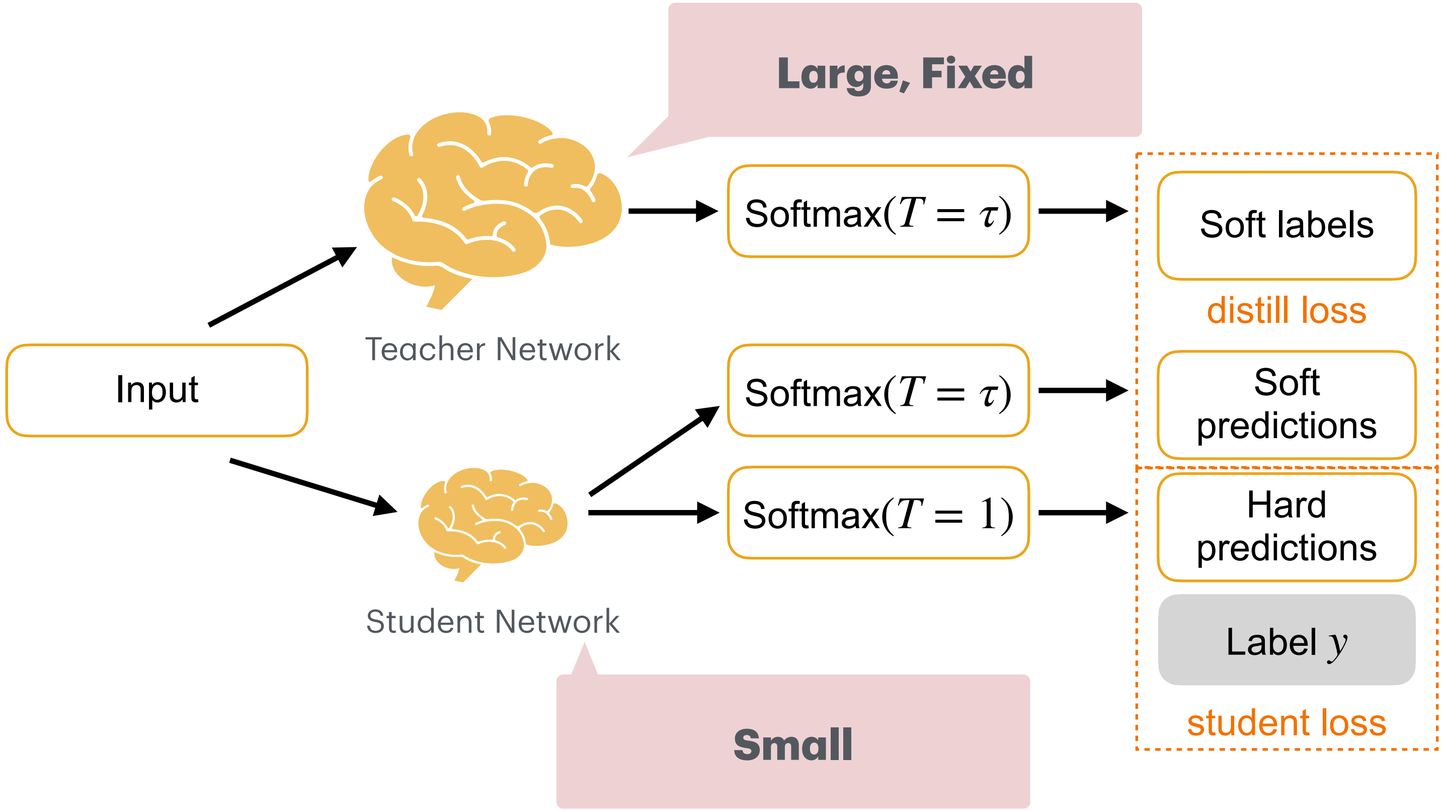
- **Output logits**
- **Gradients (attention map)**
- **Intermediate weights**
- **Sparsity patterns**
- **Intermediate features**
- **Relational information**

How to Distill?



Overview of Knowledge Distillation

Teacher model is usually larger than the student model and is fixed



- 🤔 What is the disadvantage of the fixed large teachers? Does it have to be that we need a fixed large teacher in KD?

Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2021). Knowledge distillation: A survey. International Journal of Computer Vision, 129(6), 1789-1819.

Distillation Schemes

Pre-trained

To be trained

Teacher Network

Student Network

Offline distillation

Teacher Network

Student Network

Online distillation

Teacher/Student Network

Self distillation

Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2021). Knowledge distillation: A survey. International Journal of Computer Vision, 129(6), 1789-1819.

Online Distillation

Deep Mutual Learning

- The idea of deep mutual learning: for both teacher and student networks, we want to add a distillation objective that minimizes the output distribution of the other party

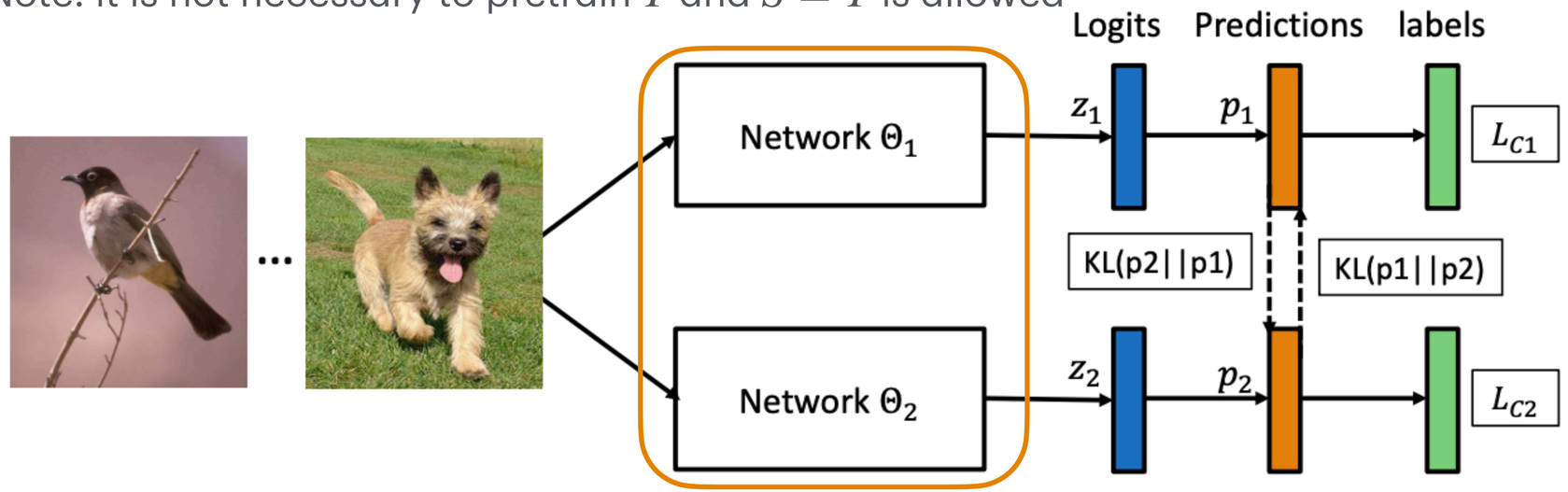
- $\mathscr{L}(S) = \text{CrossEntropy}(S(I), y) + \text{KL}(S(I), T(I))$

- $\mathscr{L}(T) = \text{CrossEntropy}(T(I), y) + \text{KL}(T(I), S(I))$

- Note: it is not necessary to pretrain T and $S = T$ is allowed

Dataset	Network Types		Independent		1 distills 2	DML	
	Net1	Net 2	Net 1	Net 2	Net 2	Net 1	Net 2
CIFAR-100	WRN-28-10	ResNet-32	78.69	68.99	69.48	78.96	70.73
	MobilNet	ResNet-32	73.65	68.99	69.12	76.13	71.10
Market-1501	Inception V1	MobileNet	65.26	46.07	49.11	65.34	52.87
	MobileNet	MobileNet	46.07	46.07	45.16	52.95	51.26

DML can improve both student (Net 2) and teacher (Net 1) models

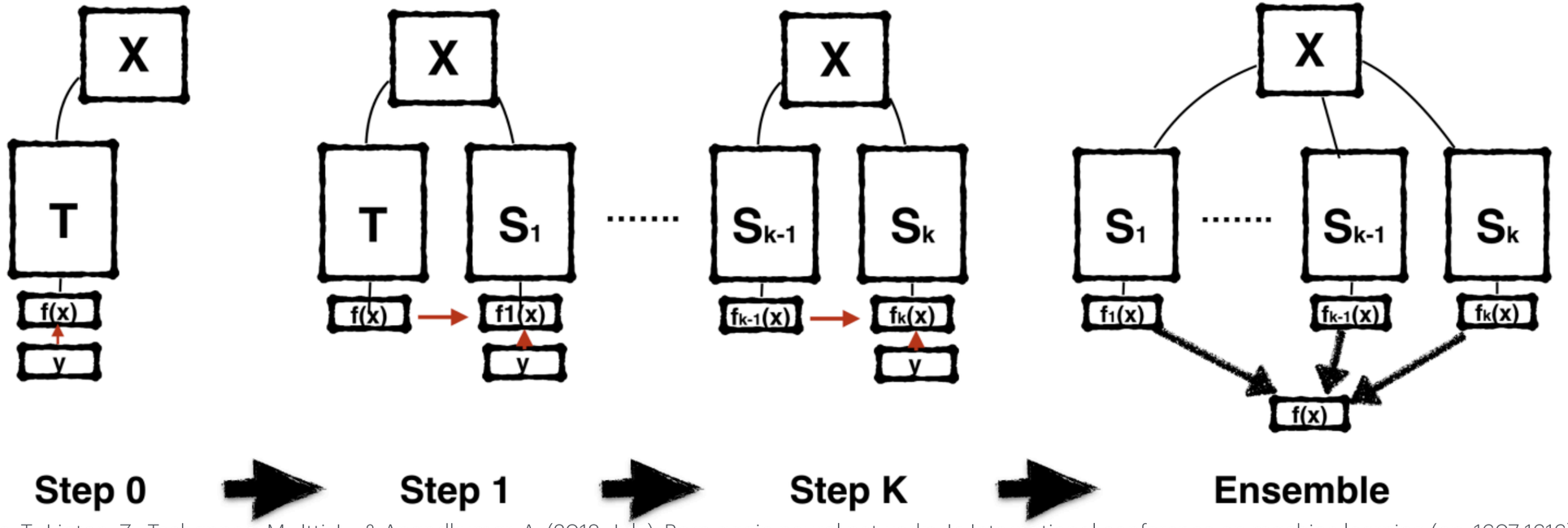


Θ_1, Θ_2 can be the same or different, and they are trained from scratch.

Zhang, Y., Xiang, T., Hospedales, T. M., & Lu, H. (2018). Deep mutual learning. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4320-4328).

Self-Distillation with Born-Again NNs

- Born-Again Network adds **iterative training stages** and uses both **classification objectives** and **distillation objectives** in subsequent stages.
- Network architecture $T = S_1 = S_2 = \ldots = S_k$ **The model sizes of the teacher and student are exactly the same.**
- Network accuracy $T < S_1 < S_2 < \ldots < S_k$
- Can alternatively **ensemble** $T, S_1, S_2, \ldots, S_k$ to get even better performance



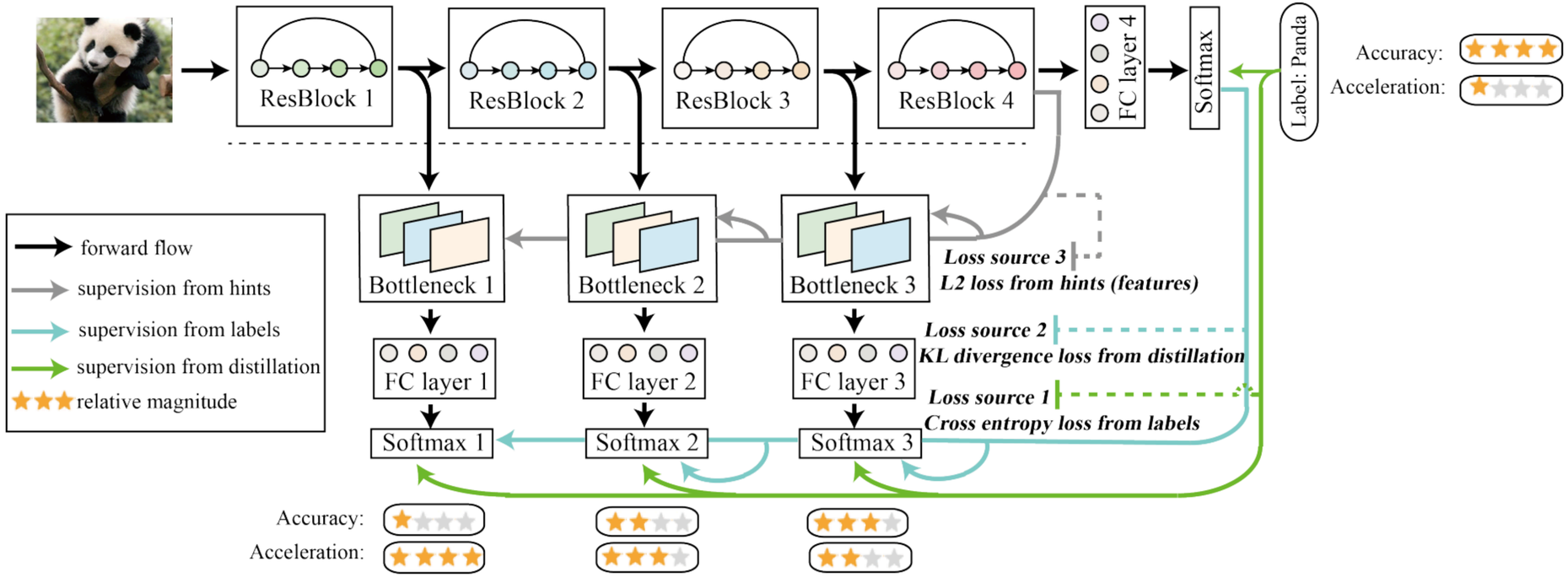
Furlanello, T., Lipton, Z., Tschannen, M., Itti, L., & Anandkumar, A. (2018, July). Born again neural networks. In International conference on machine learning (pp. 1607-1616). PMLR.

Combining Online and Self-Distillation

Be your own teacher: deep supervision + distillation

- Use deeper layers to distill shallower layers

- **Intuition**: Labels at later stages are more reliable, so the authors use them to supervise the predictions from the previous stages



Zhang, L., Song, J., Gao, A., Chen, J., Bao, C., & Ma, K. (2019). Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 3713-3722).

Combining Online and Self-Distillation

Be your own teacher: deep supervision + distillation

- Results on CIFAR100 show consistent performance improvements over the baseline

- In addition, predictions from intermediate classifiers (1/4, 2/4, 3/4) cab sometimes outperform the baseline. As such, inference efficiency can be improved.

Neural Networks	Baseline	Classifier 1/4	Classifier 2/4	Classifier3/4	Classifier 4/4	Ensemble
VGG19(BN)	64.47	63.59	67.04	68.03	67.73	68.54
ResNet18	77.09	67.85	74.57	78.23	78.64	79.67
ResNet50	77.68	68.23	74.21	75.23	80.56	81.04
ResNet101	77.98	69.45	77.29	81.17	81.23	82.03
ResNet152	79.21	68.84	78.72	81.43	81.61	82.29
ResNeXt29-8	81.29	71.15	79.00	81.48	81.51	81.90
WideResNet20-8	79.76	68.85	78.15	80.98	80.92	81.38
WideResNet44-8	79.93	72.54	81.15	81.96	82.09	82.61
WideResNet28-12	80.07	71.21	80.86	81.58	81.59	82.09
PyramidNet101-240	81.12	69.23	78.15	80.98	82.30	83.51

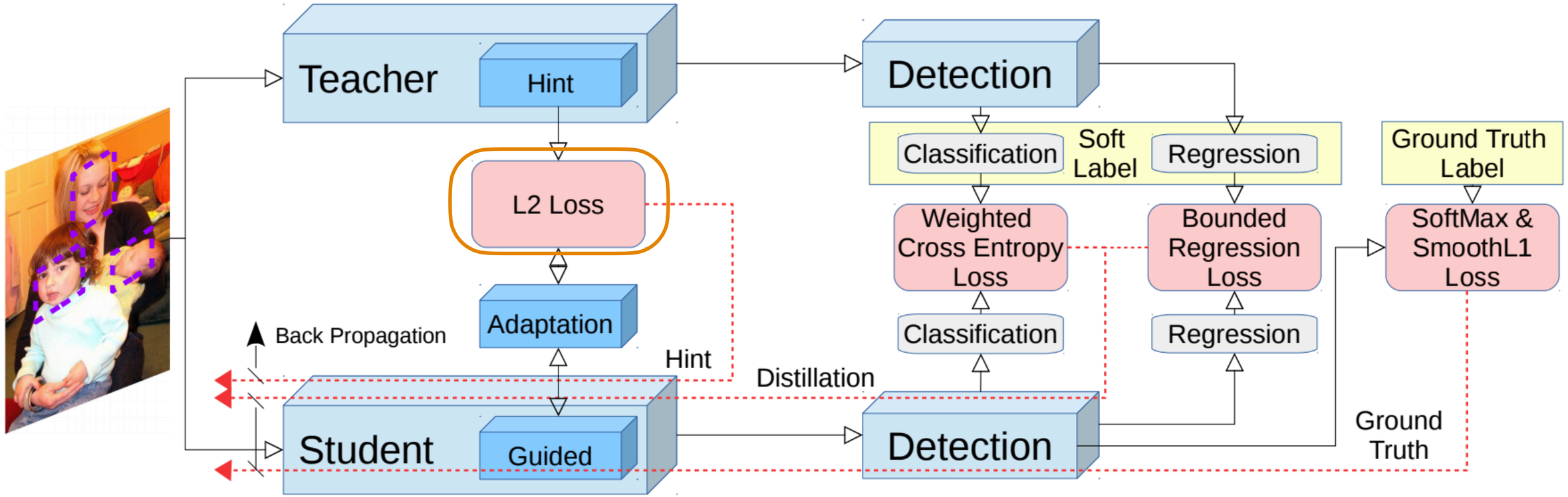
Zhang, L., Song, J., Gao, A., Chen, J., Bao, C., & Ma, K. (2019). Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 3713-3722).

Knowledge Distillation Applications

Key idea: find what knowledge to distill

KD for Object Detection

- Bounding box regression + classification



How to balance background and foreground?

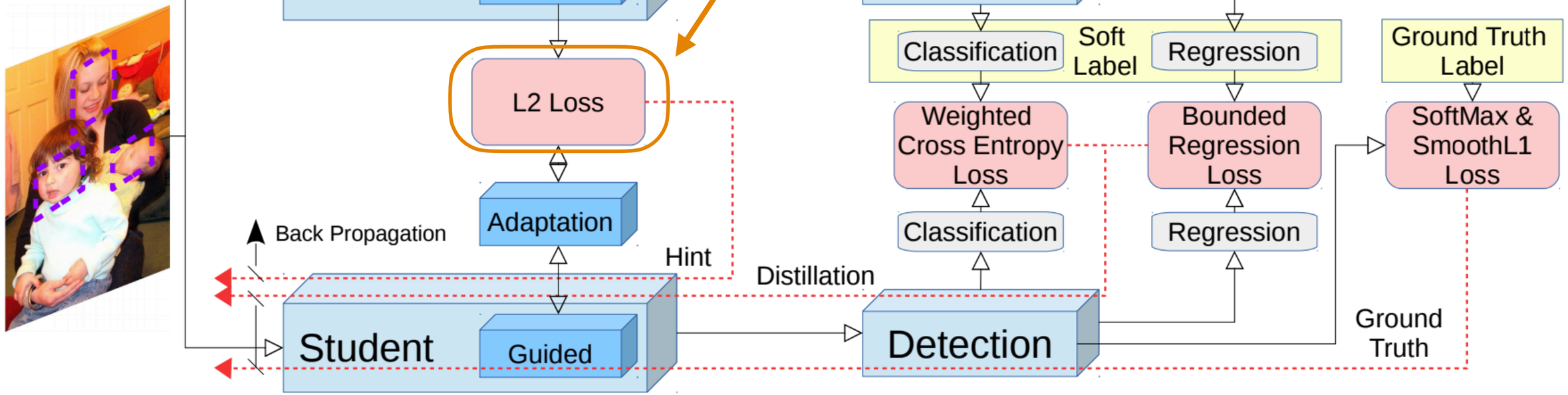
How to deal with bounding box regression?

Chen, G., Choi, W., Yu, X., Han, T., & Chandraker, M. (2017). Learning efficient object detection models with knowledge distillation. Advances in neural information processing systems, 30.

KD for Object Detection

Feature Imitation

$$L_{Hint}(V, Z) = \| V - Z \|_2^2$$



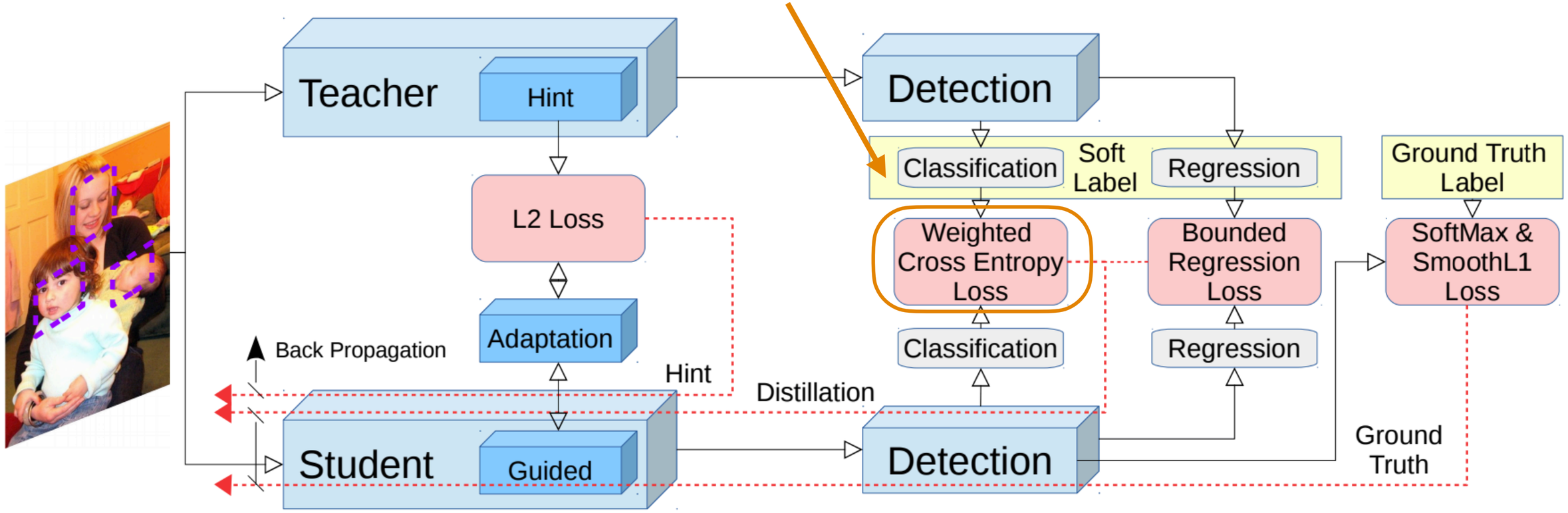
- Add a 1x1 convolution layer to match the shape

Chen, G., Choi, W., Yu, X., Han, T., & Chandraker, M. (2017). Learning efficient object detection models with knowledge distillation. Advances in neural information processing systems, 30.

KD for Object Detection

Feature Imitation

$$L_{soft}(P_s, P_t) = -\sum w_c P_t \log P_s$$



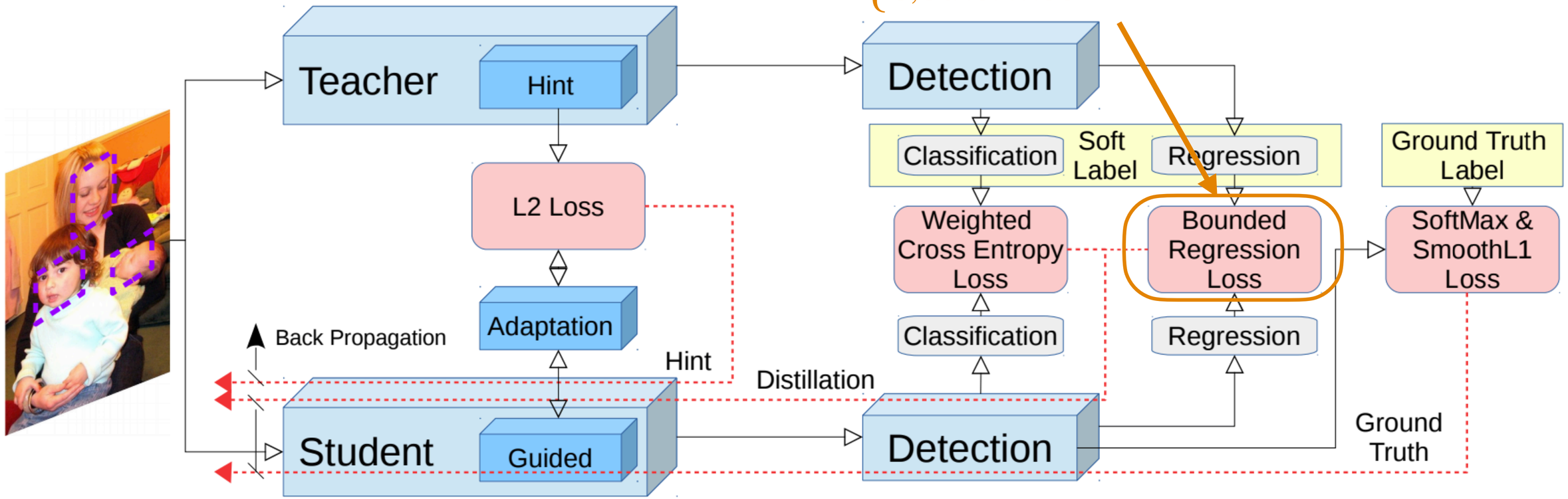
- Use different weights for foreground and background classes to handle the class imbalance problem

Chen, G., Choi, W., Yu, X., Han, T., & Chandraker, M. (2017). Learning efficient object detection models with knowledge distillation. Advances in neural information processing systems, 30.

KD for Object Detection

Feature Imitation

$$L_b(R_s, R_t, y) = \begin{cases} \| R_s - y \|_2^2, if \| R_s - y \|_2^2 + m > \| R_t - y \|_2^2 \\ 0, \text{otherwise} \end{cases}$$



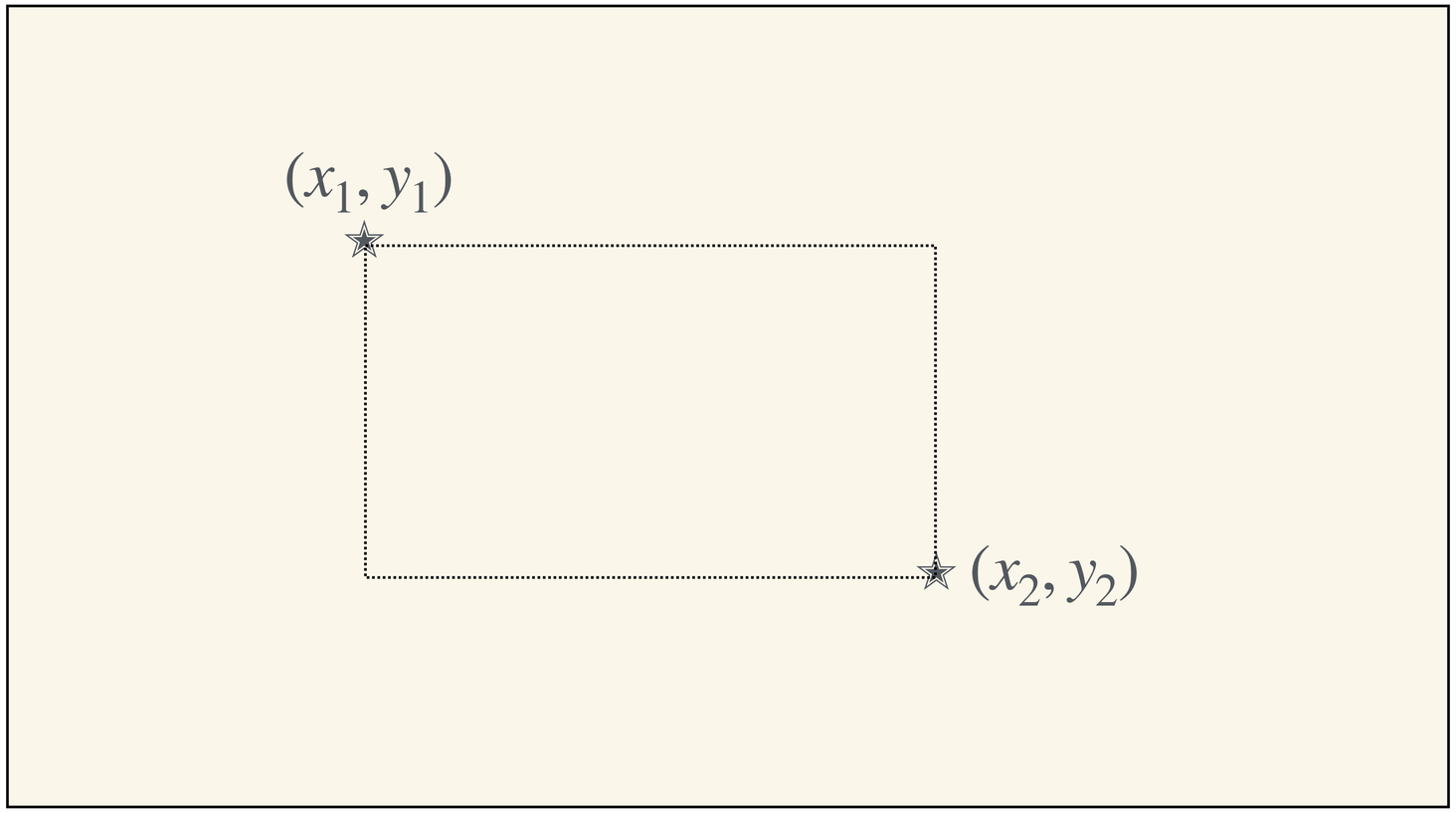
- Exploit the teacher's prediction as an upper bound for the student to achieve. Once the quality of the student surpasses that of the teacher by a certain margin, the loss becomes zero.

Chen, G., Choi, W., Yu, X., Han, T., & Chandraker, M. (2017). Learning efficient object detection models with knowledge distillation. Advances in neural information processing systems, 30.

KD for Object Detection

Convert bounding box regression to the classification problem

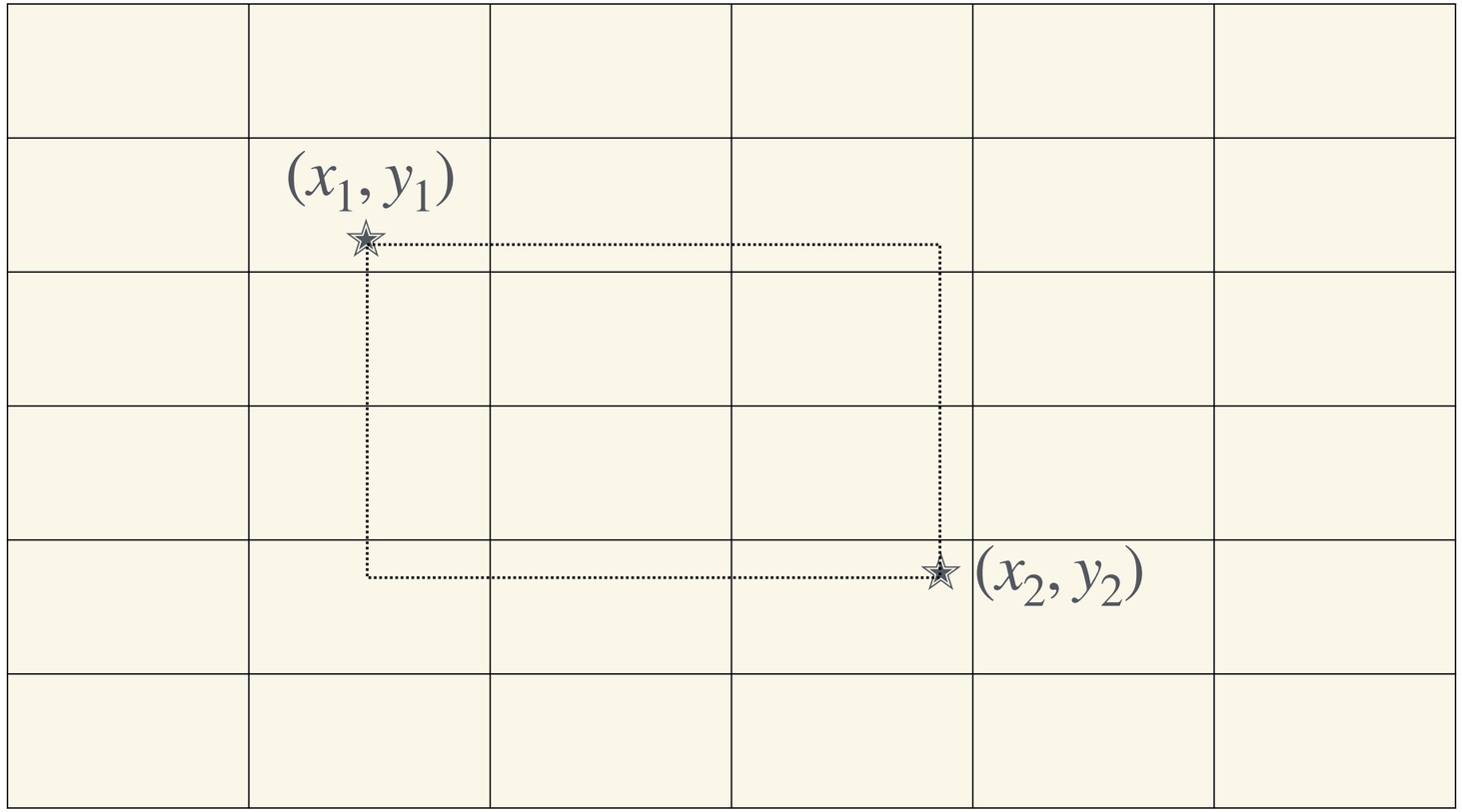
- By discretizing both the X and Y dimensions



KD for Object Detection

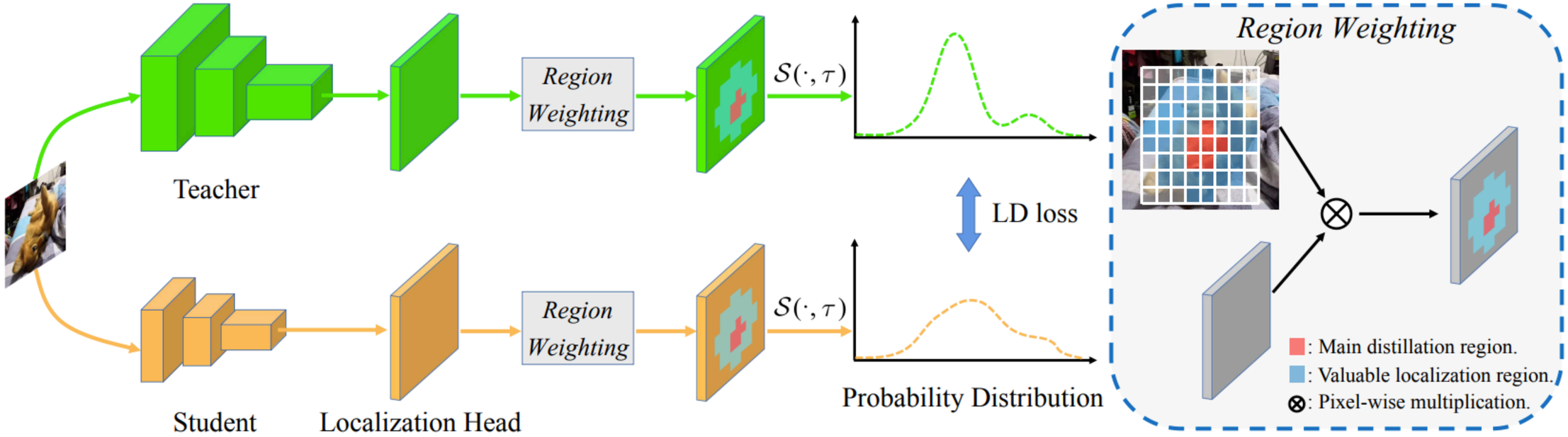
Convert bounding box regression to the classification problem

- Classification problem: x/y belongs to which bin, respectively?



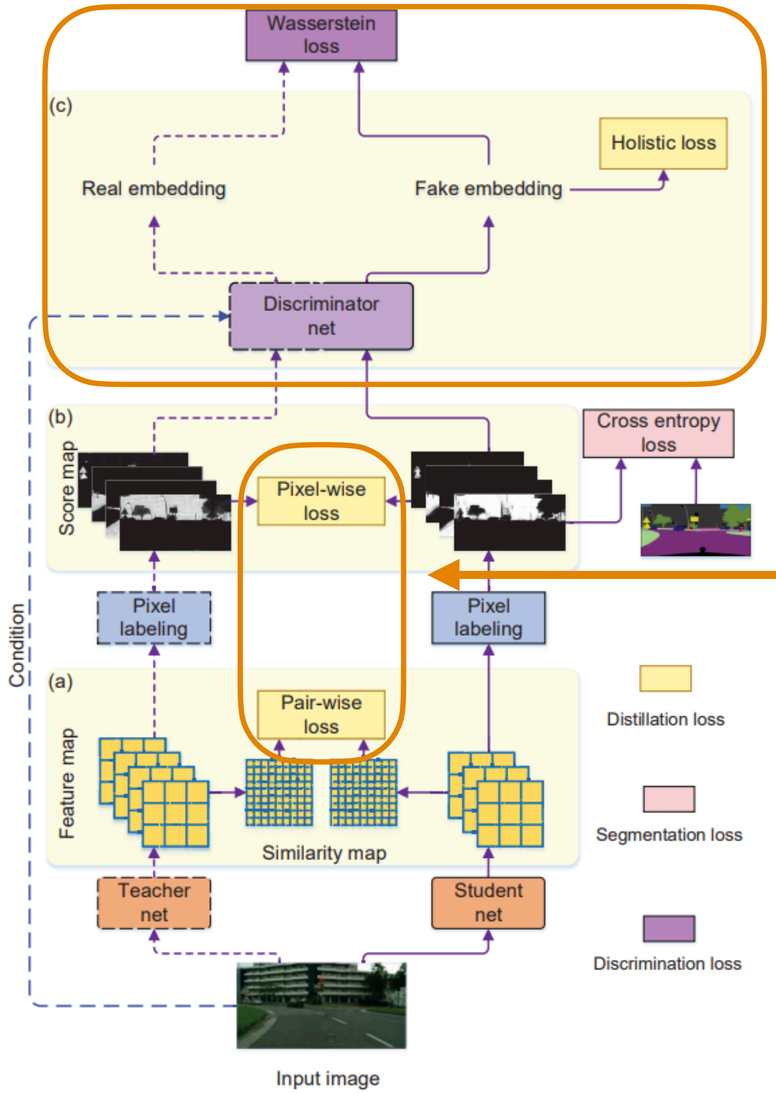
KD for Object Detection

- Localization distillation
  - Calculate the distillation loss between two probability distributions predicted by the teacher and the student



Zheng, Z., Ye, R., Wang, P., Ren, D., Zuo, W., Hou, Q., & Cheng, M. M. (2022). Localization distillation for dense object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9407-9416).

KD for Semantic Segmentation

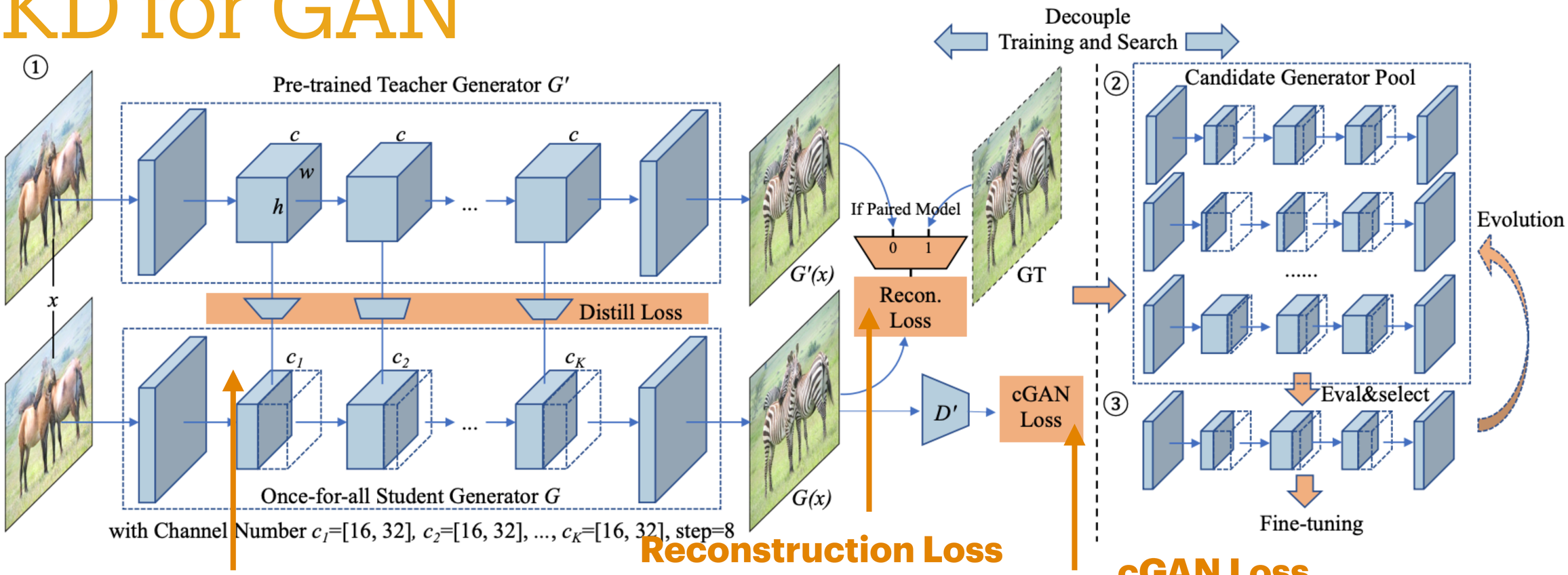


Add a discriminator network to provide adversarial loss: the student is trained to fool the discriminator network

Feature imitation similar to classification and detection

Liu, Y., Chen, K., Liu, C., Qin, Z., Luo, Z., & Wang, J. (2019). Structured knowledge distillation for semantic segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2604-2613).

KD for GAN



Distillation Loss

$$\mathscr{L}_{distill} = \sum_{k=1}^{n} \| G_k(x) - f_k(G'_k(x)) \|$$

Reconstruction Loss

$$\mathscr{L}_{recon} = \begin{cases} \| G(x) - y \| & \text{paired cGANs} \\ \| G(x) - G'(x) \| & \text{unpaired cGANs} \end{cases}$$

cGAN Loss

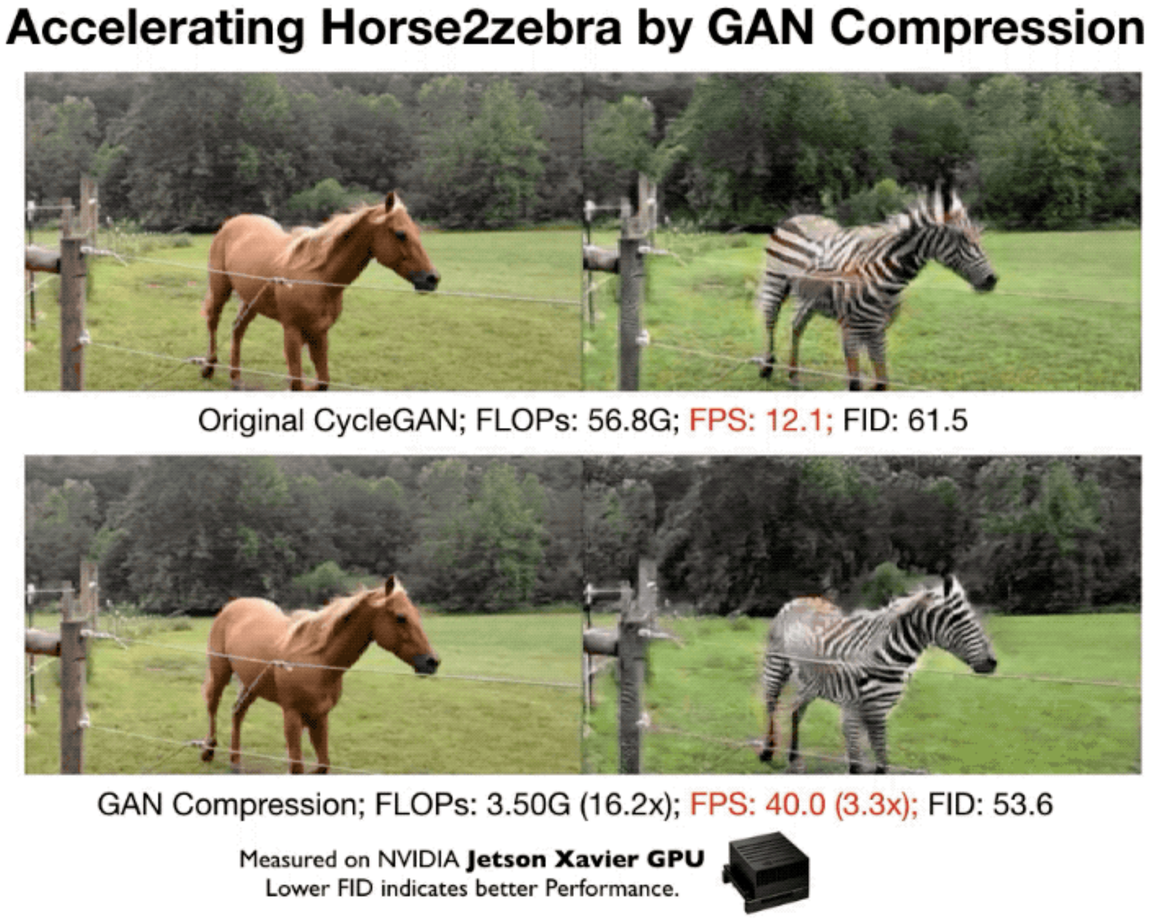
$$\mathscr{L}_{cGAN} = \mathbb{E}_{x,y}[\log D(x,y)] + \mathbb{E}_{x}[\log(1 - D(x, G(x)))]$$

Training Objective: $\mathscr{L}(x) = \mathscr{L}_{cGAN}(x) + \lambda_{recon}\mathscr{L}_{recon}(x) + \lambda_{distill}\mathscr{L}_{distill}(x)$

Li, M., Lin, J., Ding, Y., Liu, Z., Zhu, J. Y., & Han, S. (2020). Gan compression: Efficient architectures for interactive conditional gans. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5284-5294).

KD for GAN

Demo on Horse2zebra dataset



GAN Compression

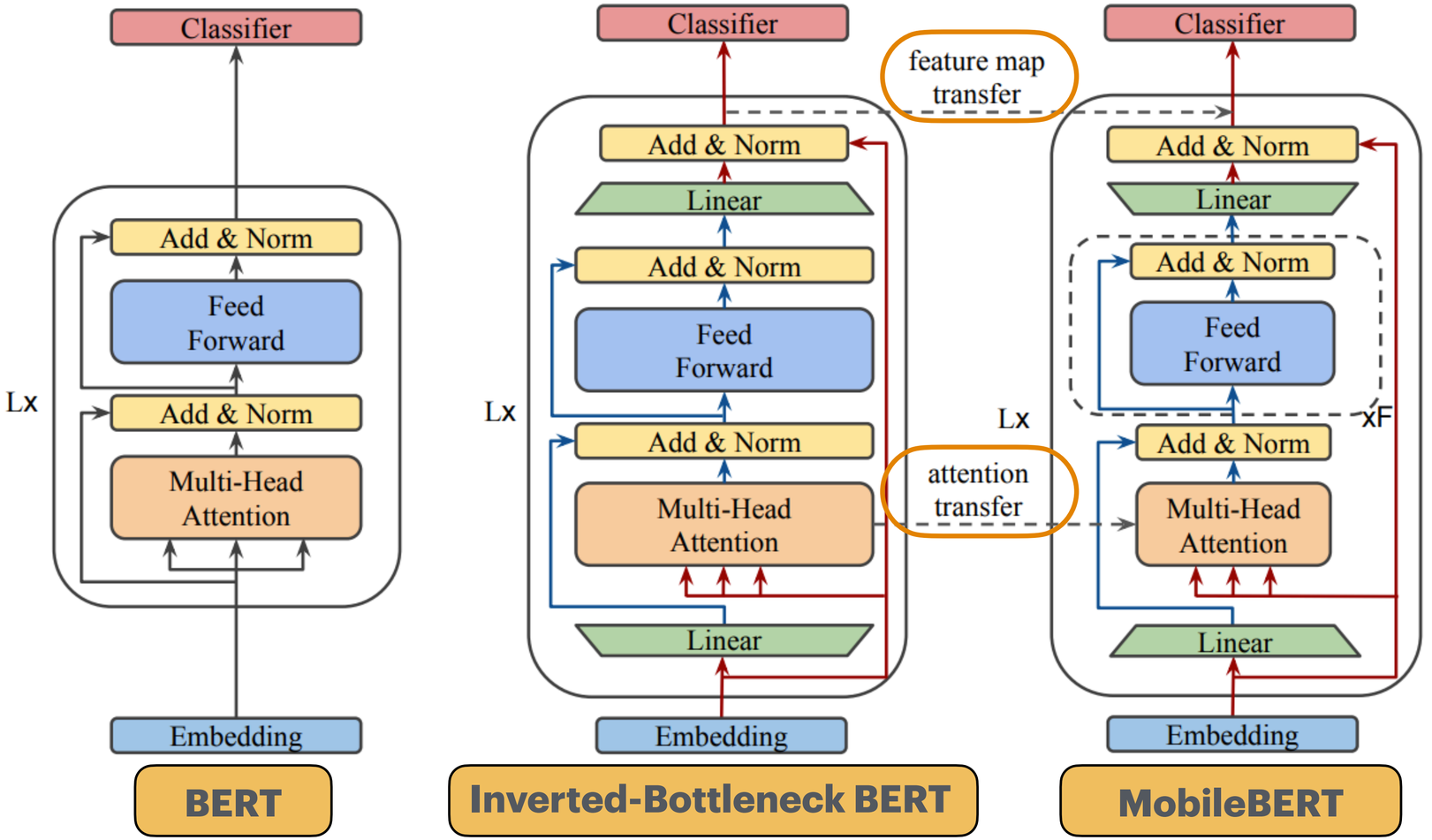
KD for GAN

Interactive image editing demo

- Try it on your own
  - https://github.com/mit-han-lab/gan-compression/tree/master/interactive_demo

KD for NLP

Attention transfer

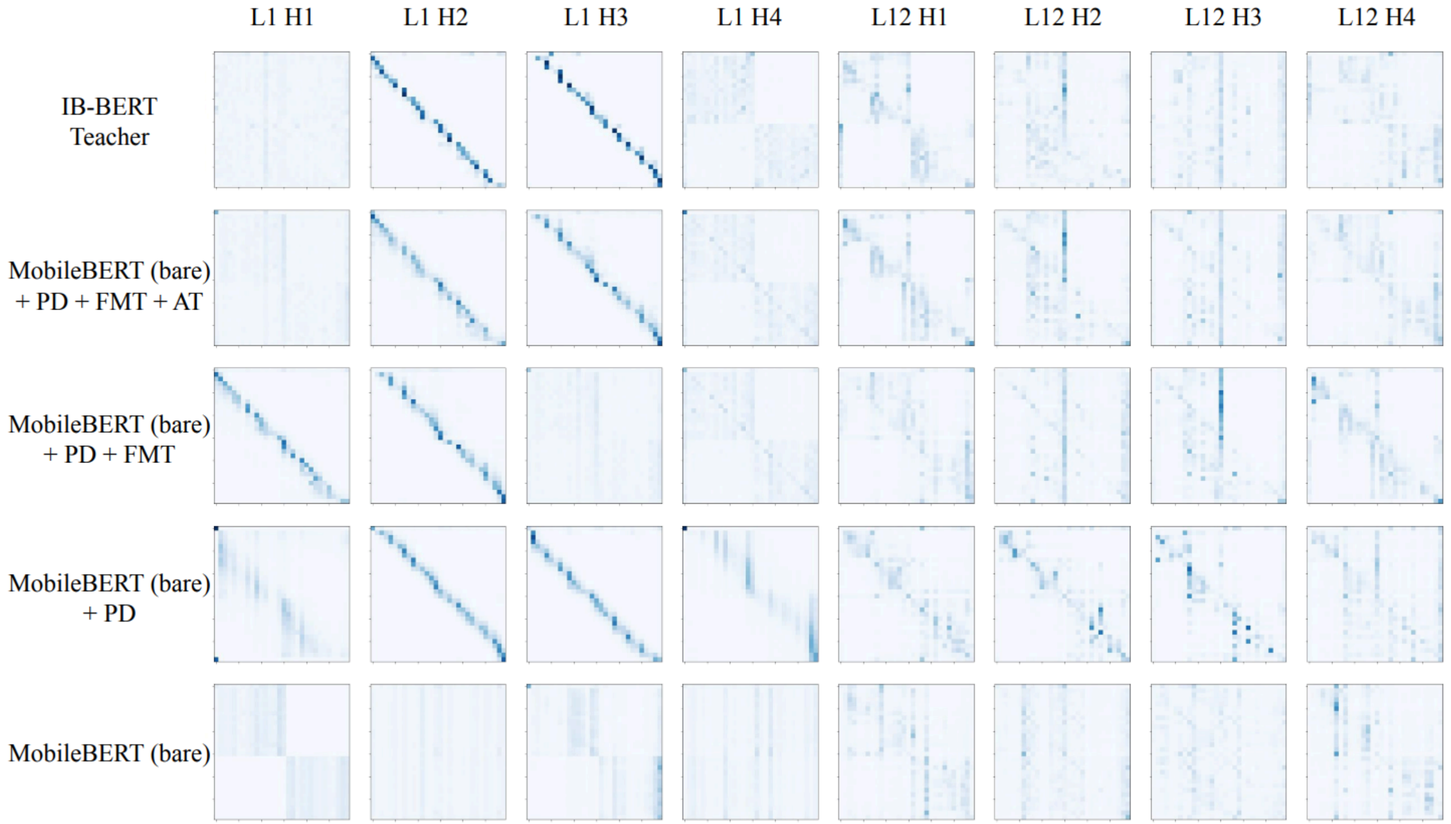


- In addition to feature imitation, the student model is trained to mimic the teacher model's attention maps

Sun, Z., Yu, H., Song, X., Liu, R., Yang, Y., & Zhou, D. (2020). Mobilebert: a compact task-agnostic bert for resource-limited devices. arXiv preprint arXiv:2004.02984.

KD for NLP

Attention transfer



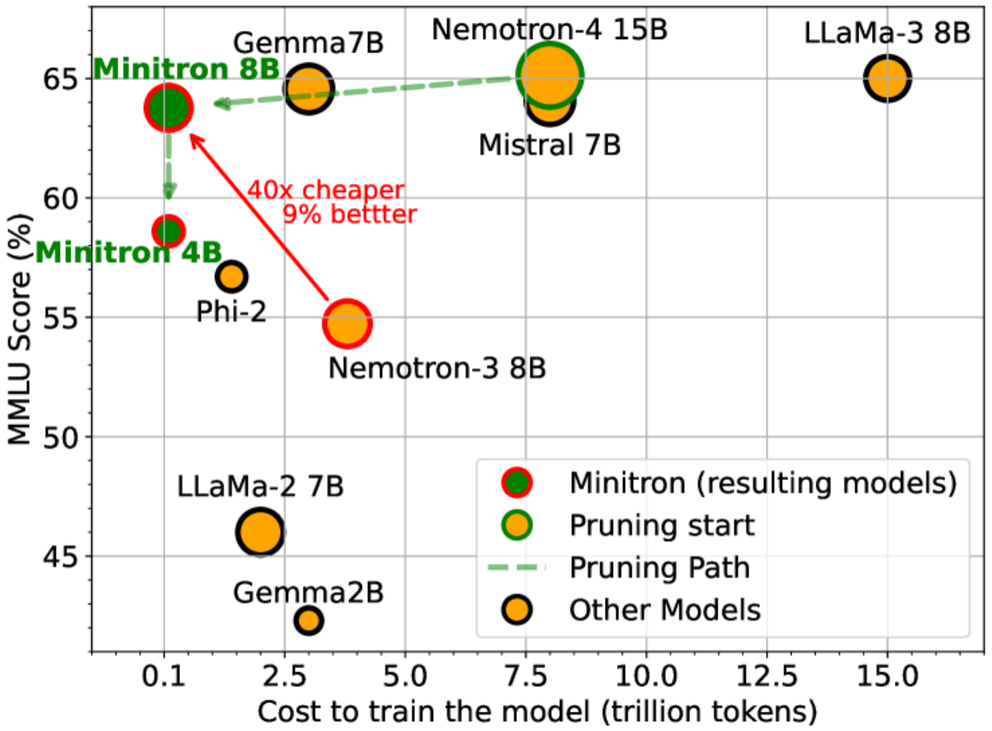
Visualization of the attention distributions in some attention heads of the IB-BERT teacher and different MobileBERT models

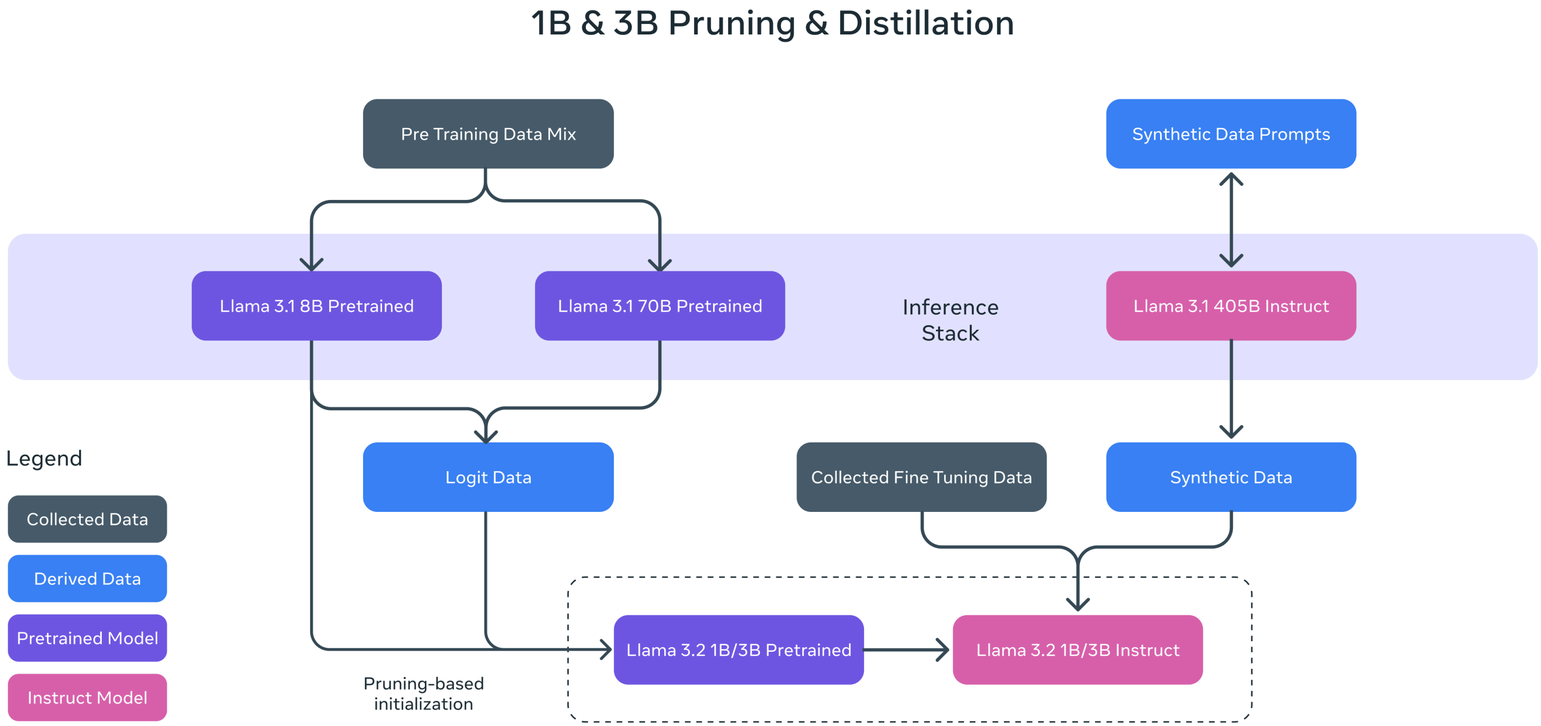
- In addition to feature imitation, the student model is trained to mimic the teacher model's attention maps

Sun, Z., Yu, H., Song, X., Liu, R., Yang, Y., & Zhou, D. (2020). Mobilebert: a compact task-agnostic bert for resource-limited devices. arXiv preprint arXiv:2004.02984.

KD for Large Language Model/Vision Language Model

Pruning and distillation become common practice to obtain small LLMs



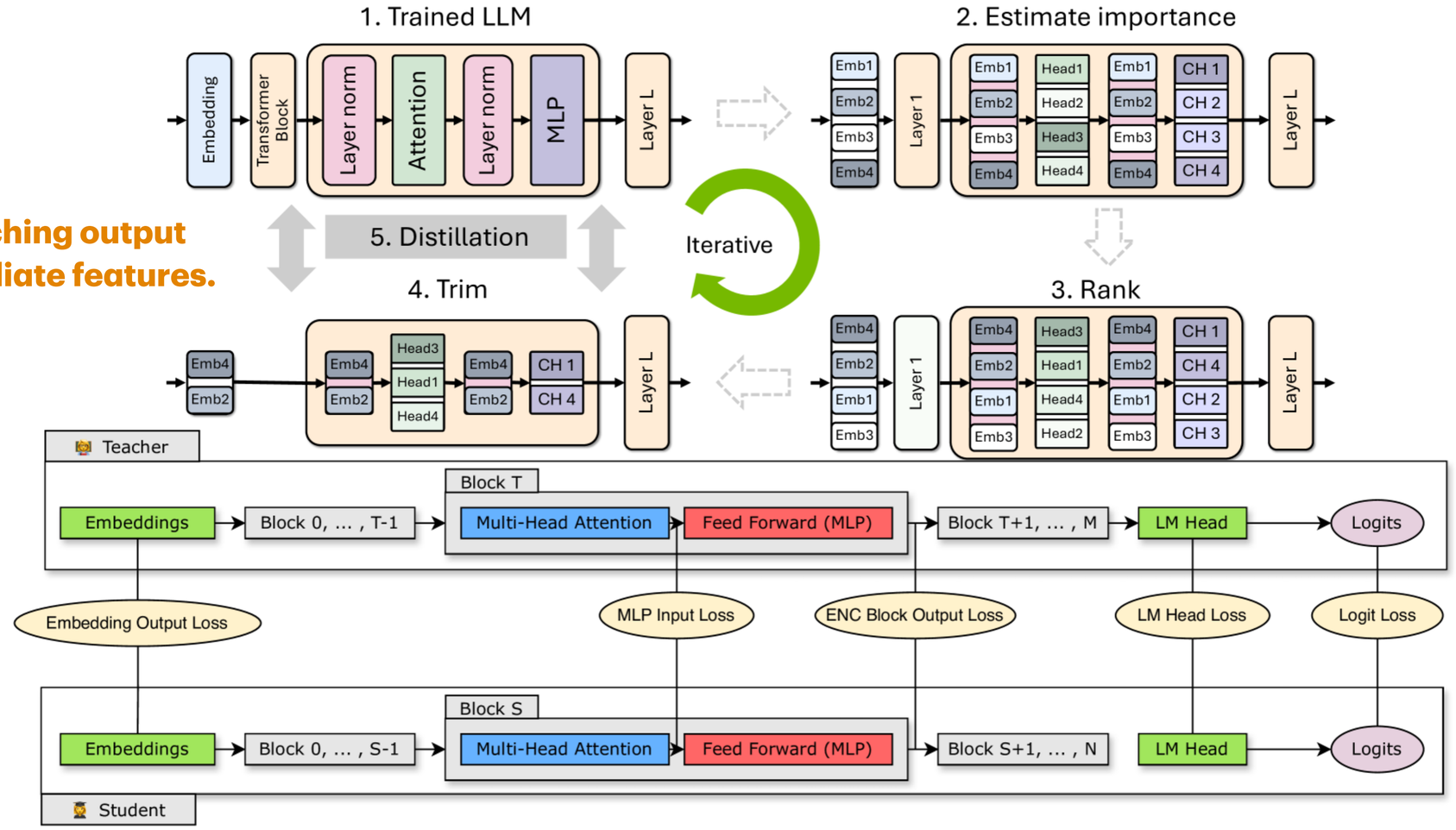


Muralidharan, S., Sreenivas, S. T., Joshi, R., Chochowski, M., Patwary, M., Shoeybi, M., ... & Molchanov, P. (2024). Compact language models via pruning and knowledge distillation. arXiv preprint arXiv:2407.14679.
Llama 3.2: Revolutionizing edge AI and vision with open, customizable models

KD for Large Language Model/Vision Language Model

Minitron applies KD during retraining after pruning the model

Distillation by matching output logits and intermediate features.



Muralidharan, S., Sreenivas, S. T., Joshi, R., Chochowski, M., Patwary, M., Shoeybi, M., ... & Molchanov, P. (2024). Compact language models via pruning and knowledge distillation. arXiv preprint arXiv:2407.14679.

Network Augmentation

An Orthogonal Way to Knowledge Distillation.

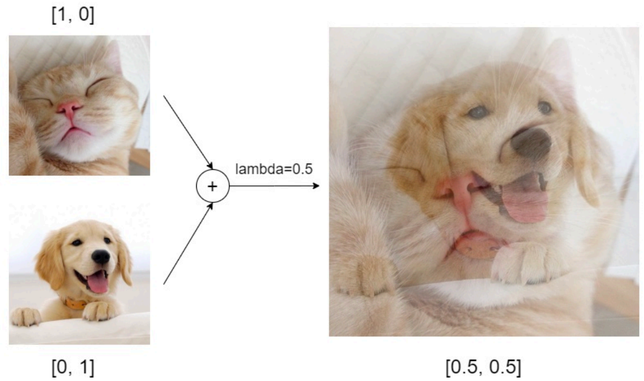
When the model is very small, how to augment the capacity?

Conventional Approach

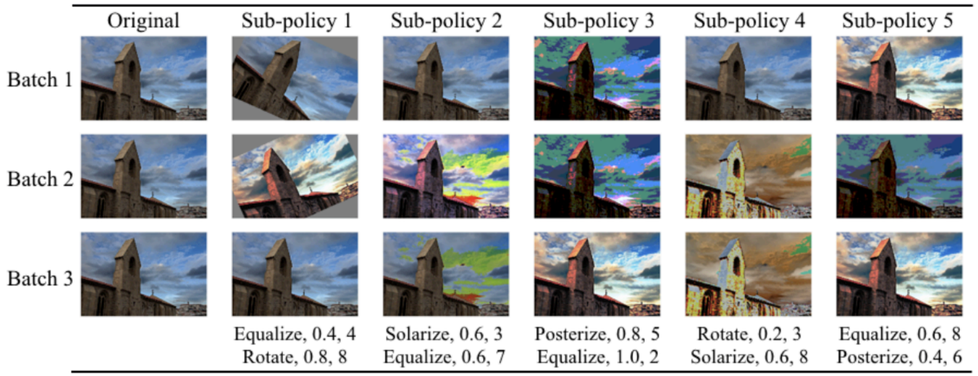
Data augmentation during training to avoid overfitting

Rescale

Cutout



Mixup



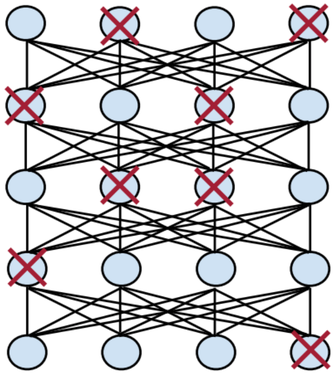
AutoAugment

🤔 What else can we do to avoid overfitting?

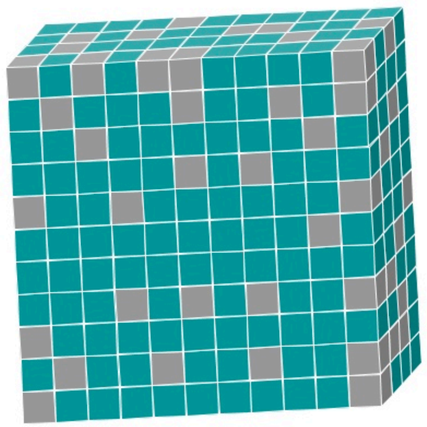
Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., & Le, Q. V. (2018). Autoaugment: Learning augmentation policies from data. arXiv preprint arXiv:1805.09501.

Conventional Approach

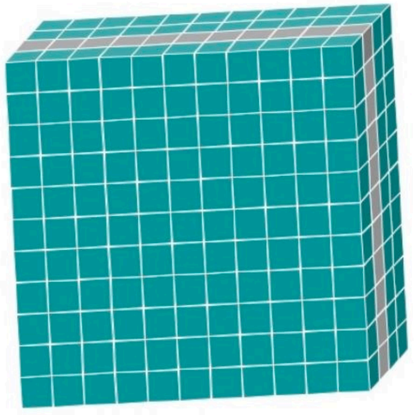
Dropout during training to avoid overfitting



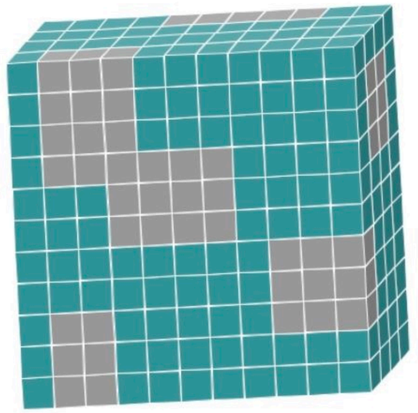
Dropout



Standard Dropout



Spatial Dropout



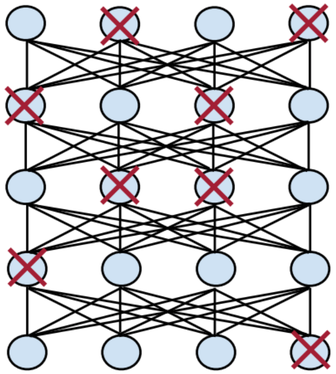
DropBlock

Ghiasi, G., Lin, T. Y., & Le, Q. V. (2018). Dropblock: A regularization method for convolutional networks. Advances in neural information processing systems, 31.

Conventional Approach

Data augmentation/dropout improves large neural network performance

Rescale

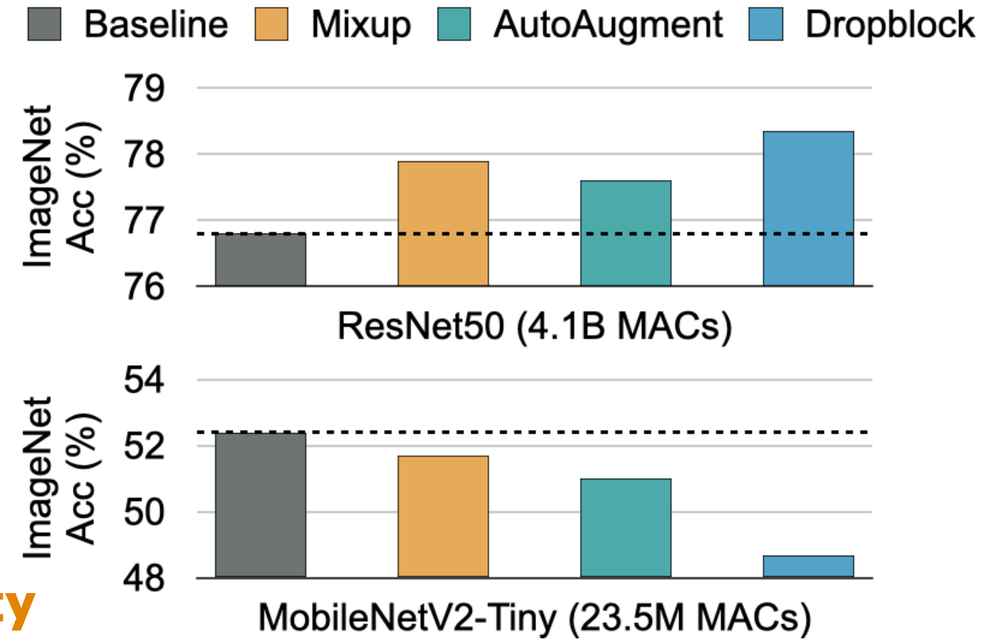


Dropout

🤩 **Improves large model's performance & prevent overfitting.**

😣 **Hurt tiny model's performance**

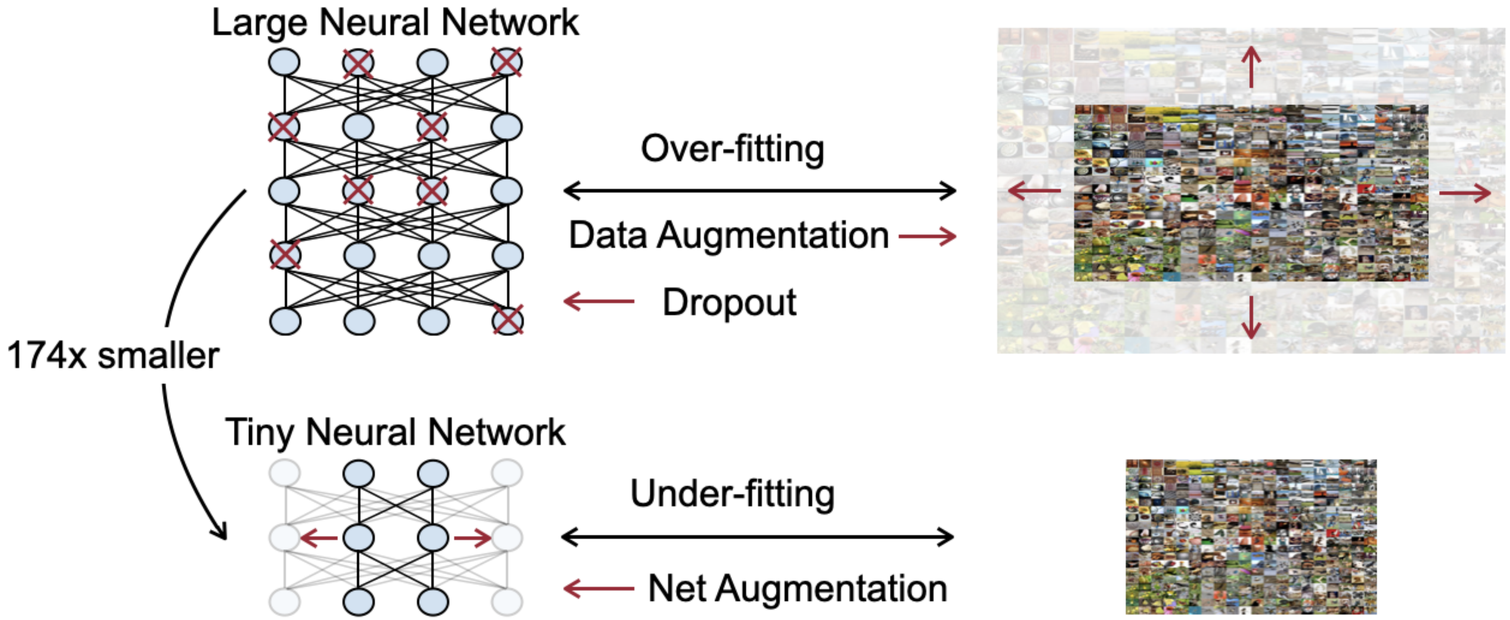
Because tiny models lack capacity

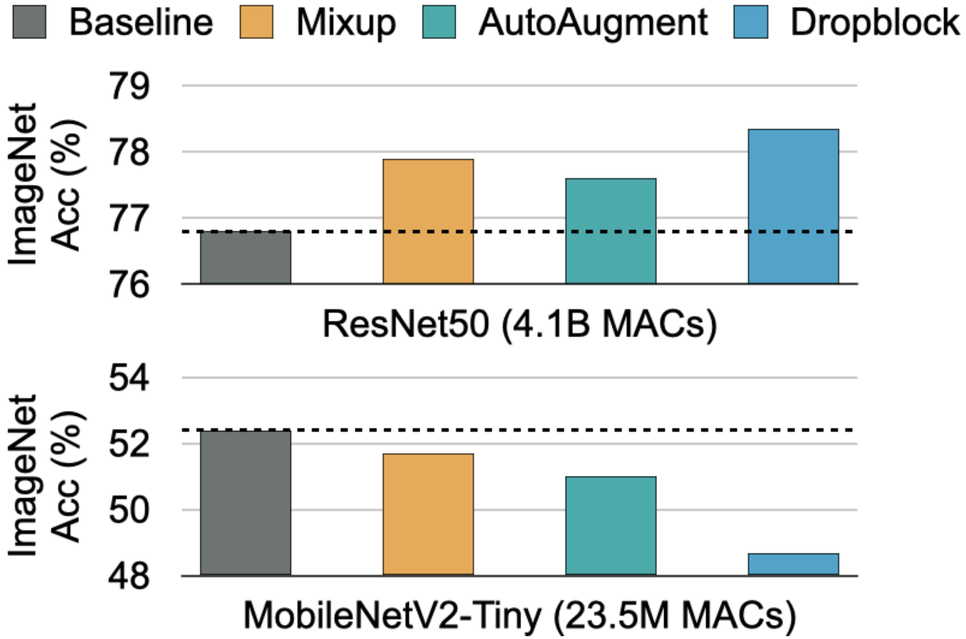


Cai, H., Gan, C., Lin, J., & Han, S. (2021). Network augmentation for tiny deep learning. arXiv preprint arXiv:2110.08890.

Conventional Approach

A tiny neural network lacks the capacity



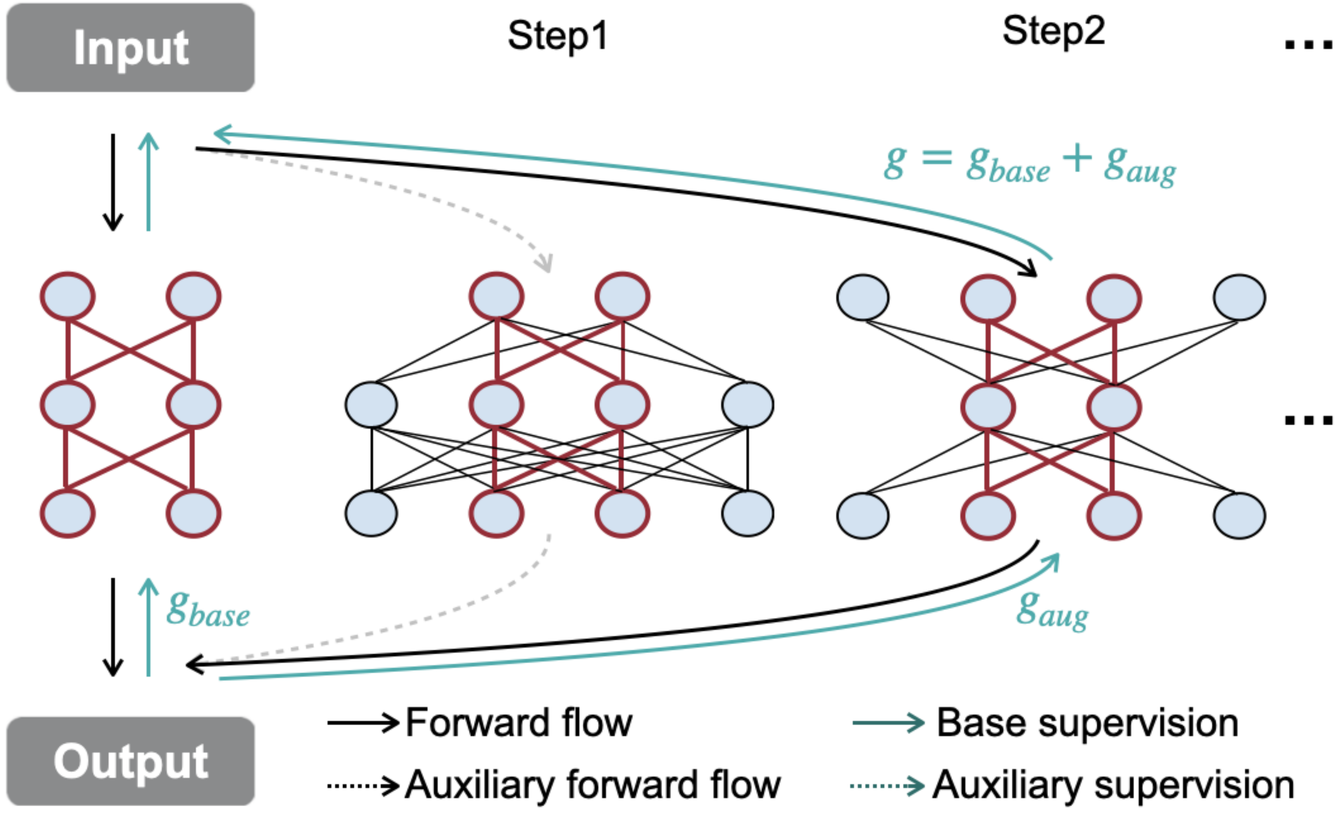


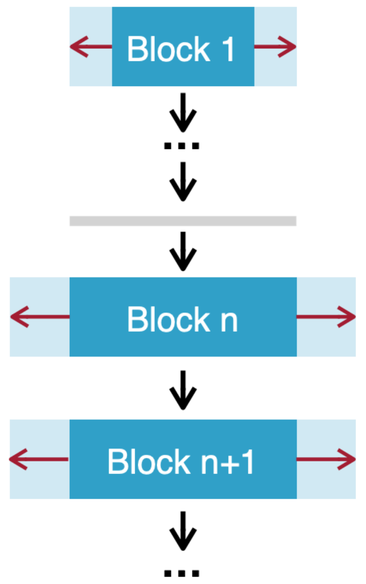
Augment the model to get extra supervision during training for tiny models

Cai, H., Gan, C., Lin, J., & Han, S. (2021). Network augmentation for tiny deep learning. arXiv preprint arXiv:2110.08890.

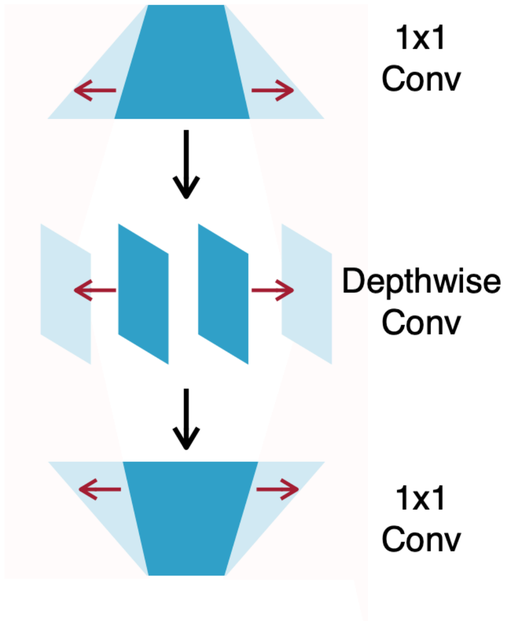
NetAug: Training Process

Augment the model to get extra supervision during training for tiny models





Augment Width Multiplier



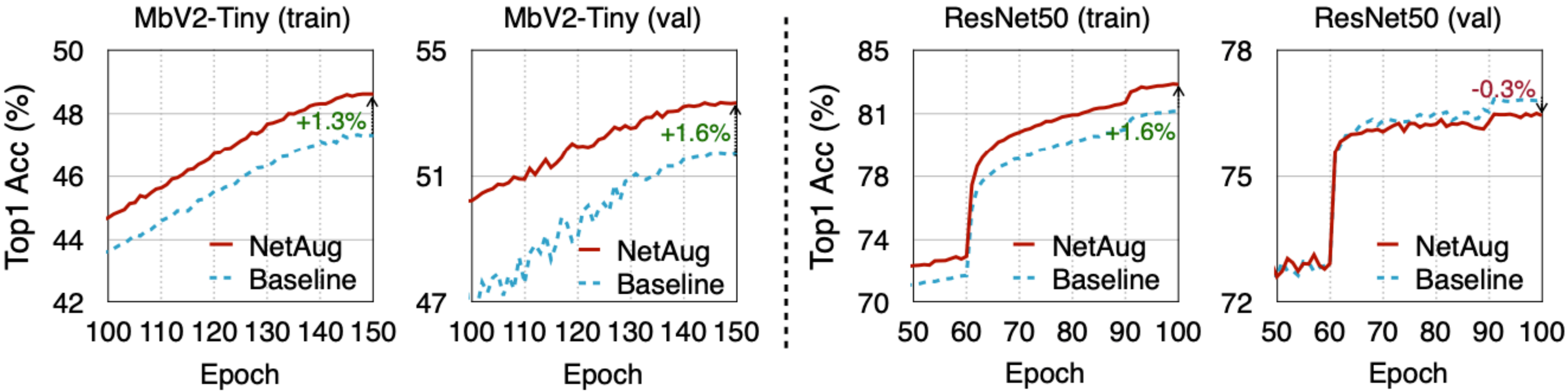
Augment Expand Ratio

$$\mathcal{L}_{\text{aug}} = \underbrace{\mathcal{L}(W_t)}_{\text{base supervision}} + \underbrace{\alpha_1\mathcal{L}([W_t, W_1]) + \cdots + \alpha_i\mathcal{L}([W_t, W_i]) + \cdots}_{\text{auxiliary supervision, working as a sub-model of augmented models}}$$

Cai, H., Gan, C., Lin, J., & Han, S. (2021). Network augmentation for tiny deep learning. arXiv preprint arXiv:2110.08890.

NetAug: Learning Curve

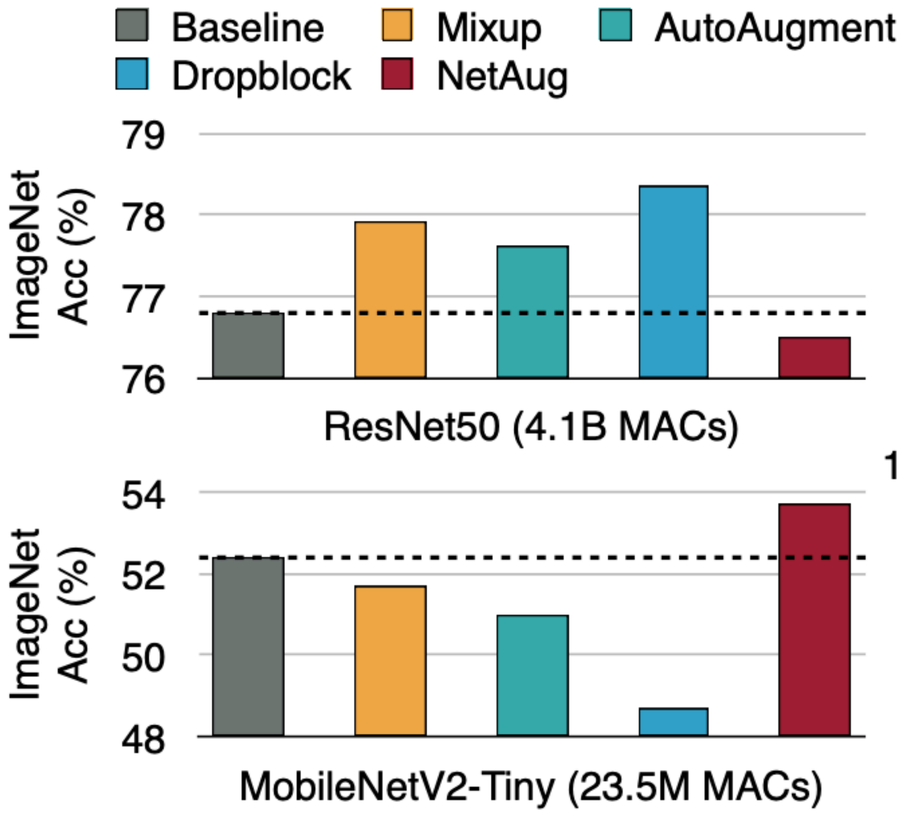
- For a tiny network, NetAug improves **both** the training accuracy and validation accuracy
- For a larger neural network, NetAug **improves** the training accuracy but **hurts** the validation accuracy

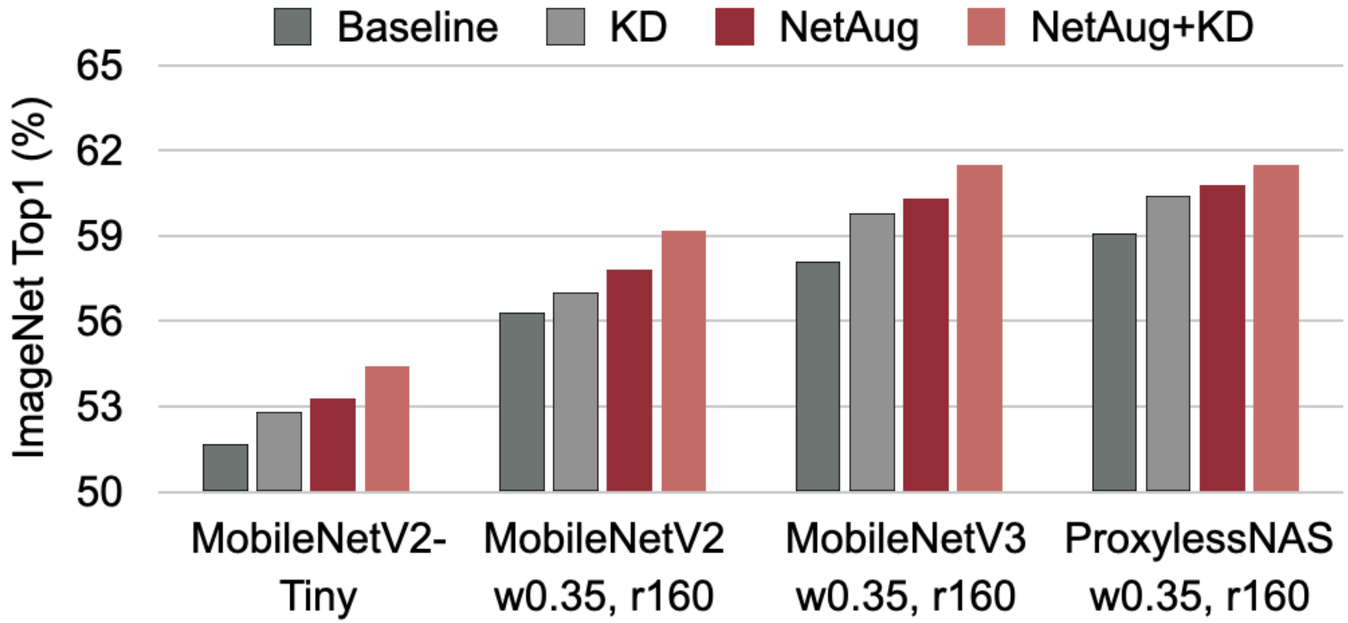


Cai, H., Gan, C., Lin, J., & Han, S. (2021). Network augmentation for tiny deep learning. arXiv preprint arXiv:2110.08890.

NetAug: Results

NetAug is orthogonal to KD

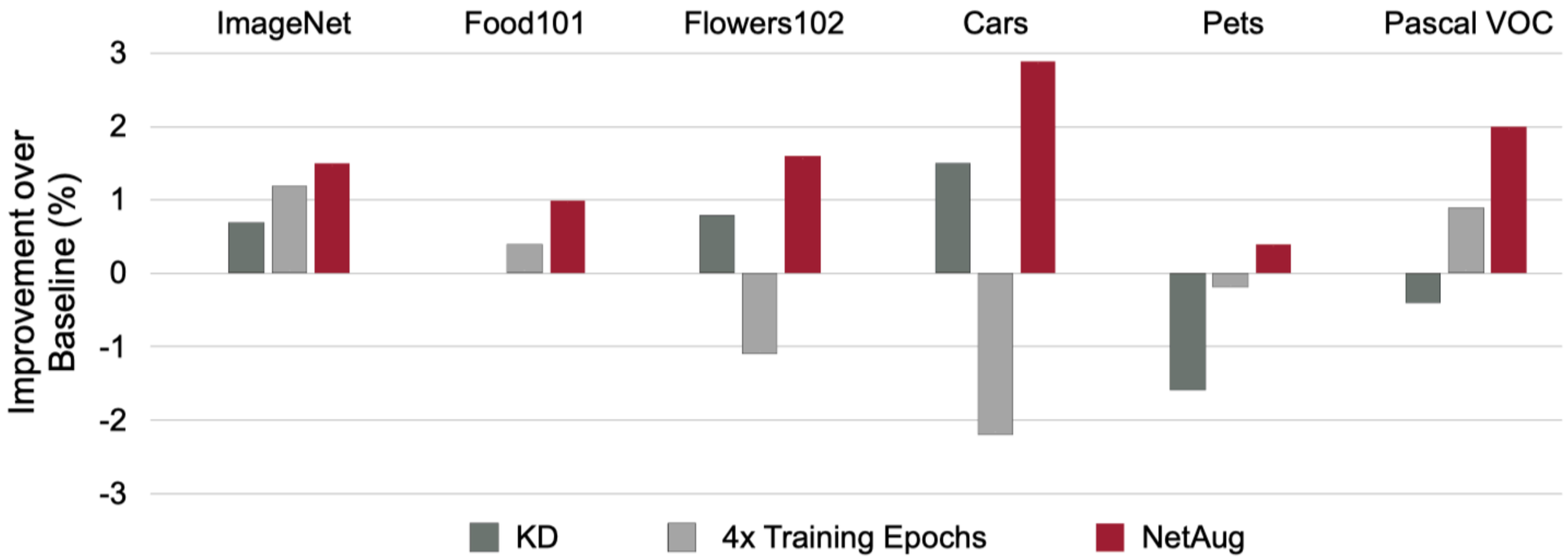




Cai, H., Gan, C., Lin, J., & Han, S. (2021). Network augmentation for tiny deep learning. arXiv preprint arXiv:2110.08890.

NetAug: Transfer Learning

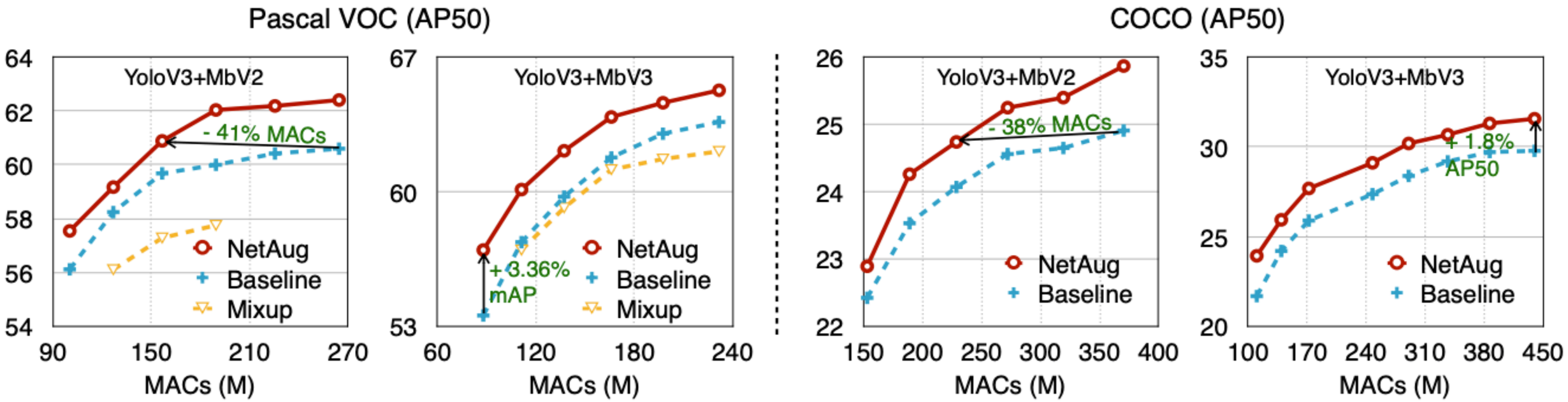
- NetAug provides better transfer learning performances than KD and 4x training schedule, though their ImageNet performances are similar



Cai, H., Gan, C., Lin, J., & Han, S. (2021). Network augmentation for tiny deep learning. arXiv preprint arXiv:2110.08890.

NetAug: Transfer to Object Detection

YoloV3 + MbV2 w0.35



- Mixup: Bag of Freebies for Training Object Detection Neural Networks

Cai, H., Gan, C., Lin, J., & Han, S. (2021). Network augmentation for tiny deep learning. arXiv preprint arXiv:2110.08890.
Zhang, Z., He, T., Zhang, H., Zhang, Z., Xie, J., & Li, M. (2019). Bag of freebies for training object detection neural networks. arXiv preprint arXiv:1902.04103.

Reference

- Cai, H., Gan, C., Lin, J., & Han, S. (2021). Network augmentation for tiny deep learning. arXiv preprint arXiv:2110.08890.
- Hinton, G. (2015). Distilling the Knowledge in a Neural Network. arXiv preprint arXiv:1503.02531.
- Ba, J., & Caruana, R. (2014). Do deep nets really need to be deep?. Advances in neural information processing systems, 27.
- Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2021). Knowledge distillation: A survey. International Journal of Computer Vision, 129(6), 1789-1819.
- Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., & Bengio, Y. (2014). Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550.
- Huang, Z., & Wang, N. (2017). Like what you like: Knowledge distill via neuron selectivity transfer. arXiv preprint arXiv:1707.01219.
- Zagoruyko, S., & Komodakis, N. (2016). Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928.
- Heo, B., Lee, M., Yun, S., & Choi, J. Y. (2019, July). Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 3779-3787).
- Yim, J., Joo, D., Bae, J., & Kim, J. (2017). A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4133-4141).
- Park, W., Kim, D., Lu, Y., & Cho, M. (2019). Relational knowledge distillation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3967-3976).
- Furlanello, T., Lipton, Z., Tschannen, M., Itti, L., & Anandkumar, A. (2018, July). Born again neural networks. In International conference on machine learning (pp. 1607-1616). PMLR.
- Zhang, Y., Xiang, T., Hospedales, T. M., & Lu, H. (2018). Deep mutual learning. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4320-4328).

Reference

- Zhang, L., Song, J., Gao, A., Chen, J., Bao, C., & Ma, K. (2019). Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 3713-3722).
- Chen, G., Choi, W., Yu, X., Han, T., & Chandraker, M. (2017). Learning efficient object detection models with knowledge distillation. Advances in neural information processing systems, 30.
- Zheng, Z., Ye, R., Wang, P., Ren, D., Zuo, W., Hou, Q., & Cheng, M. M. (2022). Localization distillation for dense object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9407-9416).
- Liu, Y., Chen, K., Liu, C., Qin, Z., Luo, Z., & Wang, J. (2019). Structured knowledge distillation for semantic segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2604-2613).
- Li, M., Lin, J., Ding, Y., Liu, Z., Zhu, J. Y., & Han, S. (2020). Gan compression: Efficient architectures for interactive conditional gans. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5284-5294).
- Sun, Z., Yu, H., Song, X., Liu, R., Yang, Y., & Zhou, D. (2020). Mobilebert: a compact task-agnostic bert for resource-limited devices. arXiv preprint arXiv:2004.02984.
- Muralidharan, S., Sreenivas, S. T., Joshi, R., Chochowski, M., Patwary, M., Shoeybi, M., ... & Molchanov, P. (2024). Compact language models via pruning and knowledge distillation. arXiv preprint arXiv:2407.14679.
- Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., & Le, Q. V. (2018). Autoaugment: Learning augmentation policies from data. arXiv preprint arXiv:1805.09501.
- Ghiasi, G., Lin, T. Y., & Le, Q. V. (2018). Dropblock: A regularization method for convolutional networks. Advances in neural information processing systems, 31.
- Zhang, Z., He, T., Zhang, H., Zhang, Z., Xie, J., & Li, M. (2019). Bag of freebies for training object detection neural networks. arXiv preprint arXiv:1902.04103.
- Knowledge Disilltation [MIT 6.5940]