Knowledge Distillation

TODO: This is in old format where i link a bunch of paper in a topic which i dont want
Knowledge distillation is the process of lossy transfer from a model to another while (most of the time) retaining the accuracy and decreases model complexity.

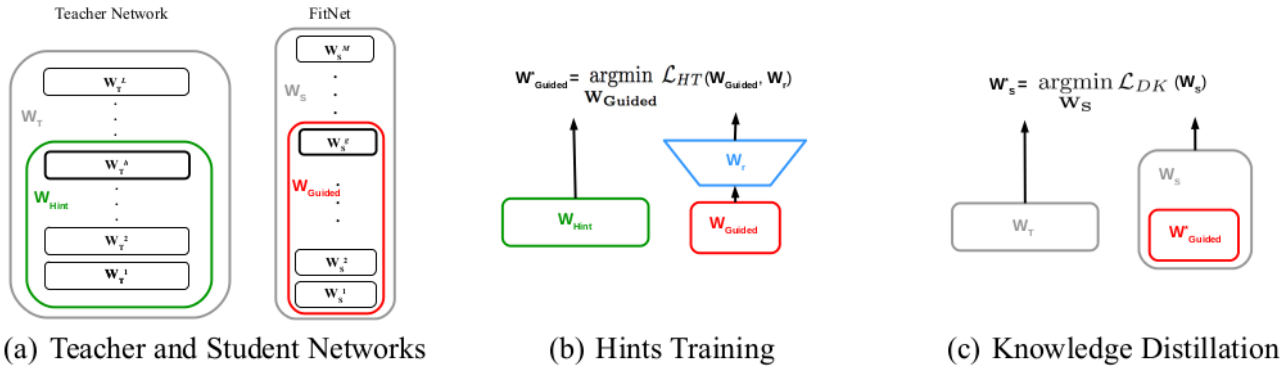
Papers

Distilling the Knowledge in a Neural Network

- Distillation by training the student model with soft targets produced from teacher model
  - Soft targets are produced by increasing "Temperature" of the softmax function, Higher "Temperature" produces softer distributions
  - $q_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}$
- $L = \frac{T^2 CE(soft) + CE(hard)}{2}$

FitNets: Hints for Thin Deep Nets

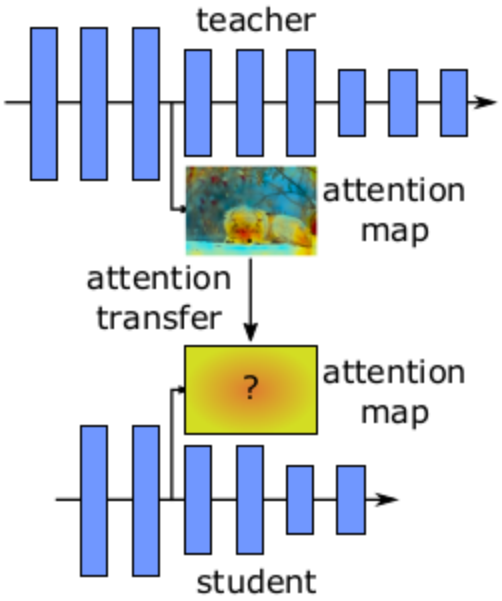
- Uses intermediate-level hints from the teacher hidden layers to guide the training process of the student

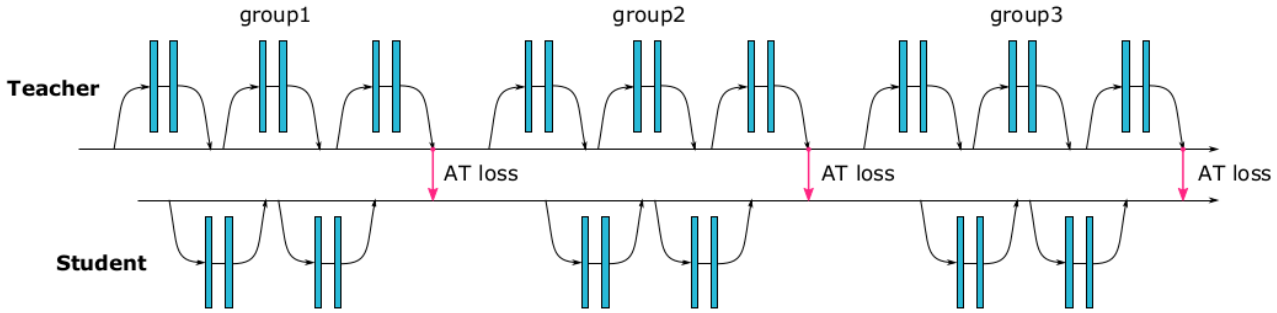


(a) Teacher and Student Networks (b) Hints Training (c) Knowledge Distillation

Paying More Attention to Attention: Improving the Performance of CNN via Attention Transfer

- Uses Attention map (activation based and gradient based) to transfer knowledge





A Gift From Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning

- Flow of solution procedure (FSP) captures relationship between feature maps from different layers. It is calculated using the inner product of features, similar to Gram matrix used for texture representation

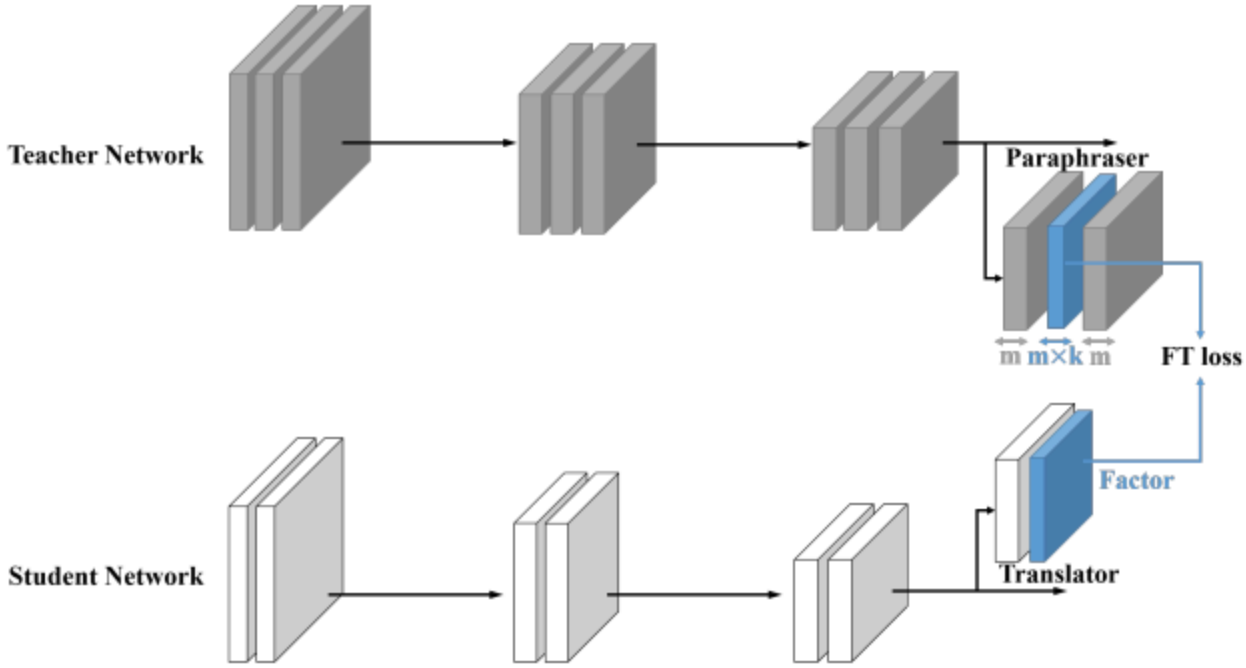
Learning Deep Representations with Probabilistic Knowledge Transfer

- Probabilistic KT(PKT): matching probability distributions of data within their respective feature spaces

Paraphrasing Complex Network: Network Compression via Factor Transfer

- Factor Transfer(FT): trains the student to replicate "factors", which are paraphrased knowledge extracted from the teacher network

- Paraphraser extracts "teacher factors" from the teacher using convolutional layers
- Translator extracts "student factors" and mimics the teacher factors



- Limitations:
  - additional parameters

Knowledge Transfer via Distillation of Activation Boundaries Formed by Hidden Neurons

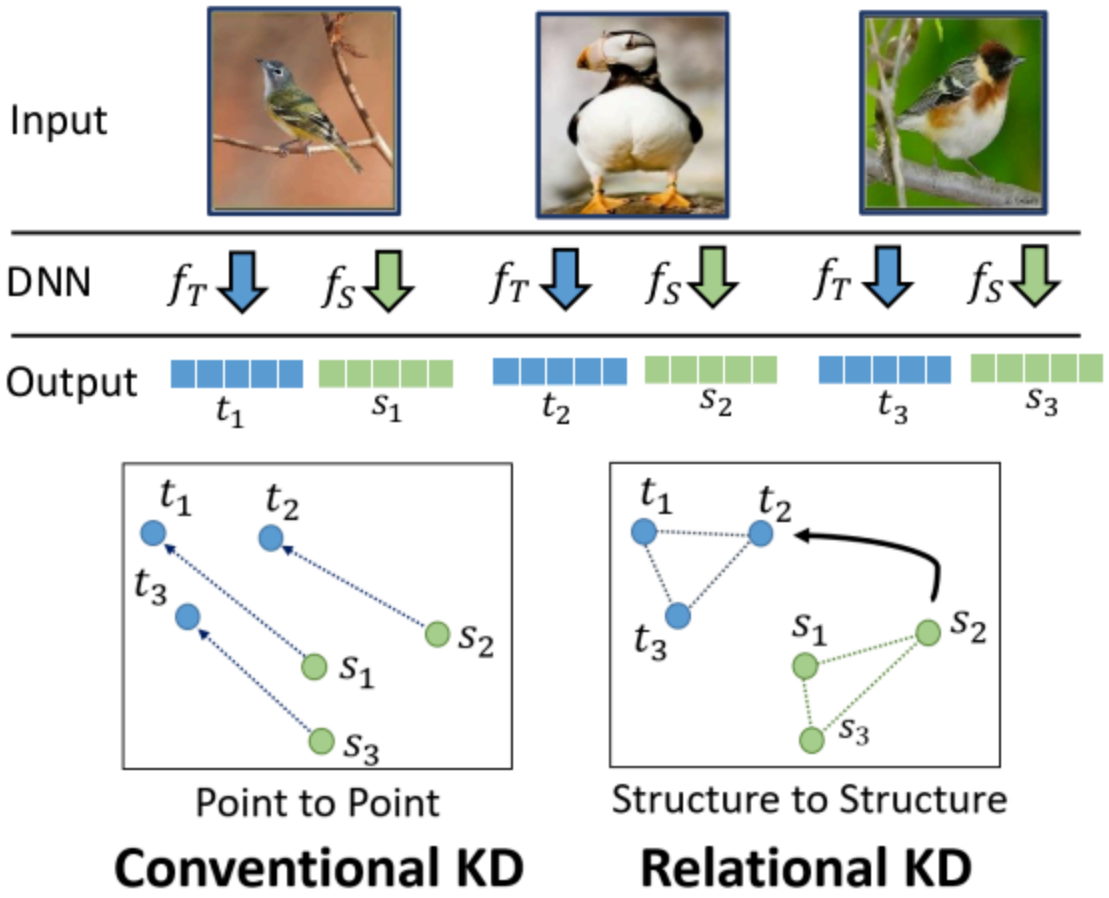
- Distilling the activation boundaries formed by the hidden neurons in a teacher network to a student network
- Activation Transfer Loss: minimizes the difference in neuron activations between teacher and student, regardless of the magnitude of the response
- Piecewise Differentiable Loss: To overcome the non-differentiability of the activation transfer loss, a hinge loss-inspired alternative loss is proposed, enabling gradient-based optimization

Knowledge Adaptation for Efficient Semantic Segmentation

- Distillation for Semantic Segmentation
- Efficiency-Accuracy Trade-off: Reducing feature map resolution via subsampling increases efficiency but compromises accuracy
- Knowledge Translation: A pre-trained autoencoder rephrases the teacher's knowledge into a compact representation

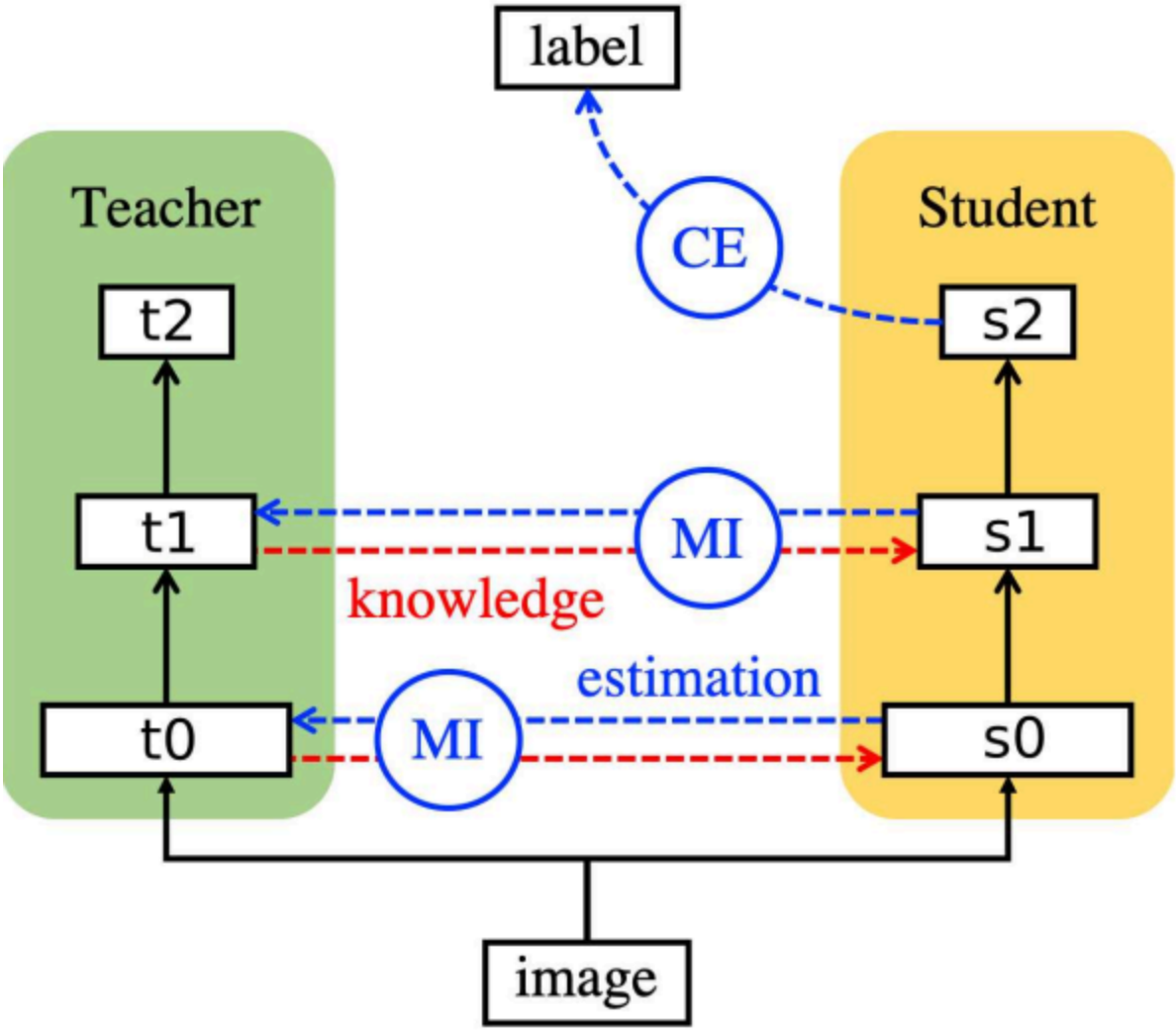
Relational Knowledge Distillation*

- Relational Knowldge Distillation (RKD): Transfer mutual relations of data examples instead with distance-wise and angle-wise distillation losses that penalize structural differences



Variational Information Distillation for Knowledge Transfer

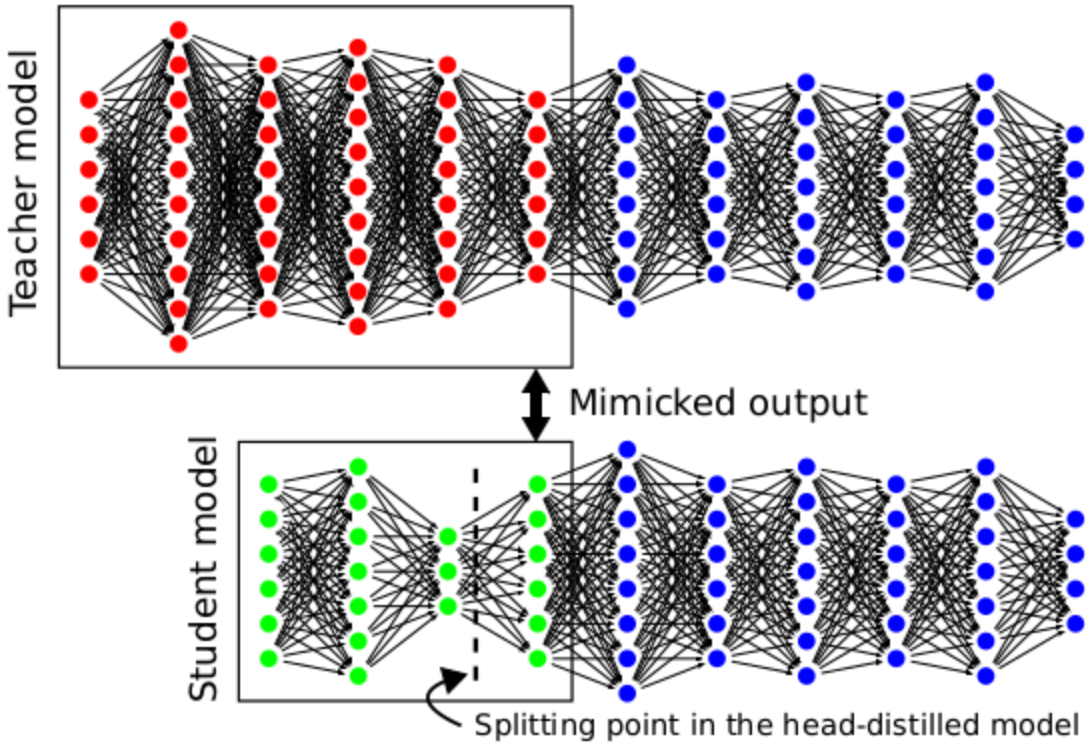
- Minimizing CE loss while retaining high mutual Information (MI) with the teacher network.
  - MI is maximized by learning to estimate the distribution of the activations in the teacher network



[Distilled Split Deep Neural Networks for Edge-Assisted Real-Time Systems]

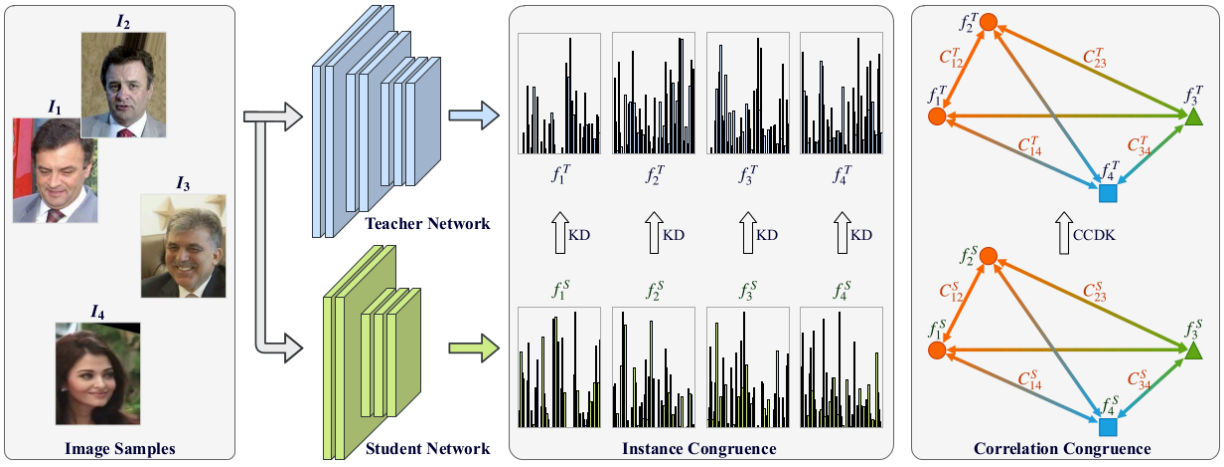
- Split DNN models into head and tail models, where the two sections are deployed at the mobile device and edge servers

- The head model is distilled from the teacher model to mimic its output



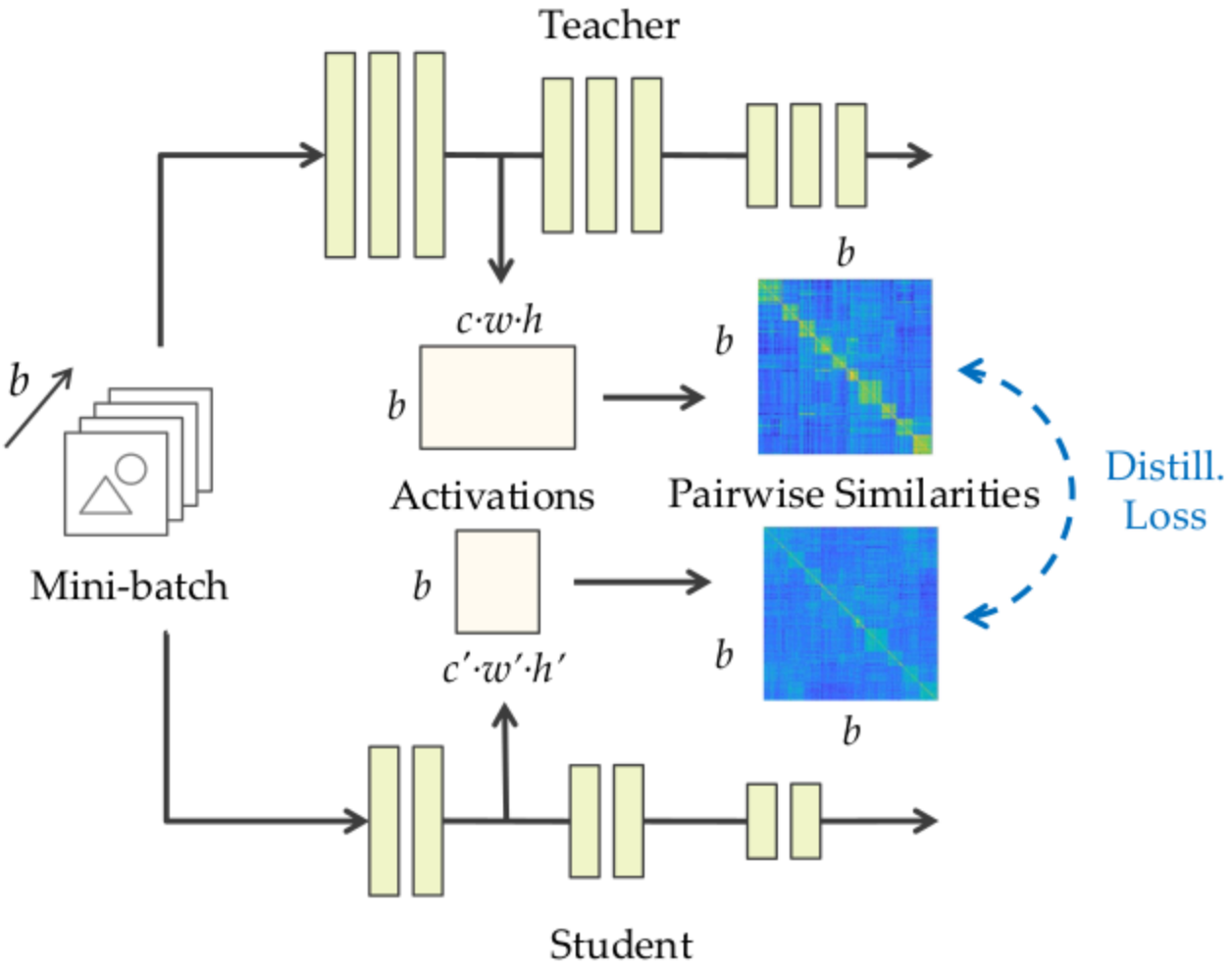
Correlation Congruence for Knowledge Distillation

- CCKD: Transfers not only the instance-level in formation but also the correlation between instances
  - by introducing a correlation congruence constraint, it aims to match the correlation matrix of the student network's feature representations with that of the teacher netowork



Similarity-Preserving Knowledge Distillation

- Preserve the pairwise similarities derived from the activations of the teacher and student networks for a given mini-batch of inputs



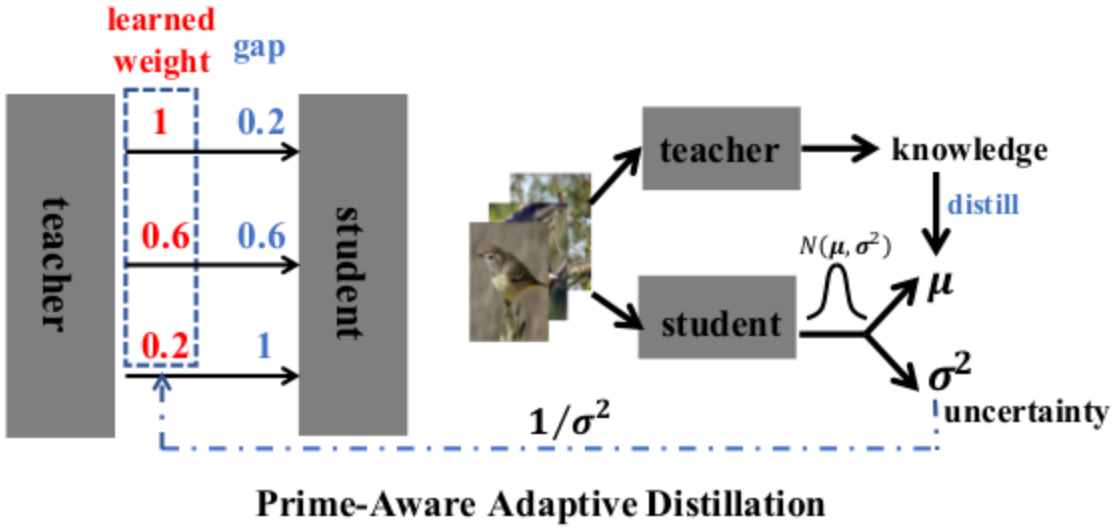
Contrastive Representation Distillation

- Transferring structural knowledge between teacher and student models by maximizing the mutual information between their representations
  - maximizing a lower bound on the mutual information between teacher and student representations

Prime-Aware Adaptive Distillation

- Learn from the obvious
- Hard samples have detrimental effect on the training of the student model
- Modeling knowledge distillation with data uncertainty (the student model Generates a mean a.k.a. the prediction and a variance, the lower the variance the higher the confidence hence

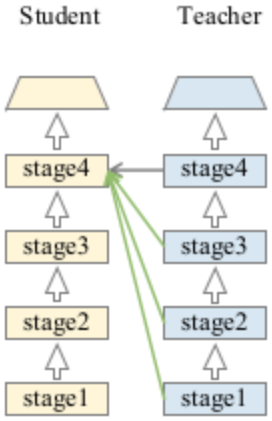
the prime samples are assigned larger learned weights)



Prime-Aware Adaptive Distillation

Distilling Knowledge via Knowledge Review

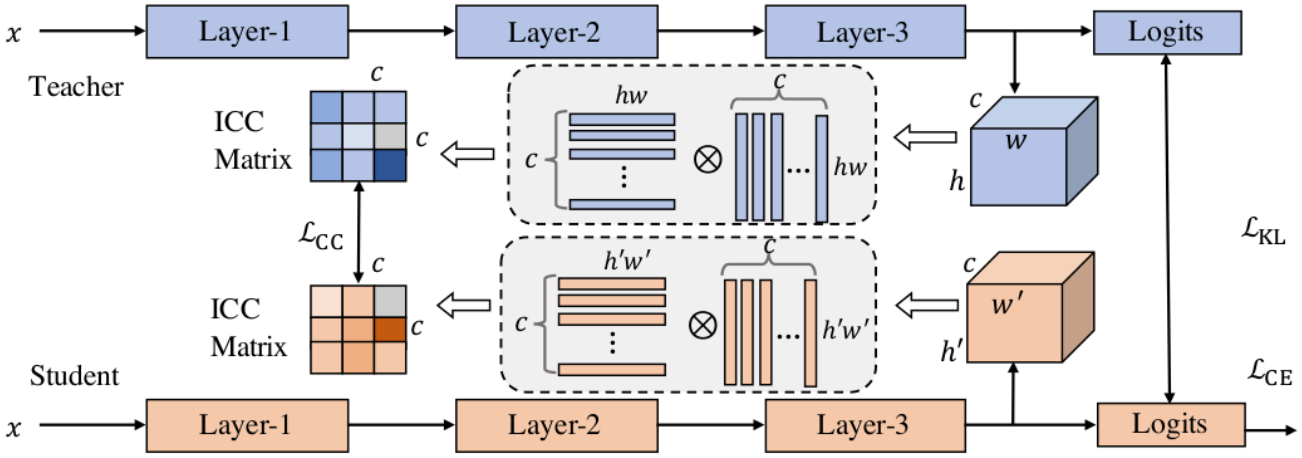
- Use low-level features in the teacher network to supervise deeper features for the student



- Attention Based Fusion (ABF): Dynamically aggregates features from different levels using attention maps, enabling adaptive integration of diverse information
- Hierarchical Context Loss (HCL): Employs spatial pyramid pooling to distill knowledge at different levels of contextual abstraction, facilitating more comprehensive knowledge transfer

Exploring Inter-Channel Correlation for Diversity-Preserved Knowledge Distillation

- Student mimics teacher's ICC matrix



Knowledge Distillation from A Stronger Teacher

- Proposes to preserve the relations between predictions rather than matching the exact values
  - Uses Pearson correlation coefficient as a metric to measure the relationship between the teacher and student predictions
  - preserving the preference by the teacher, instead of recovering the absolute values accurately

Understanding the Role of the Projector in Knowledge Distillation

- Projection layer, often used for dimension matching, plays a much more crucial role in KD
- While larger projectors can theoretically encode more information, they tend to decorrelate input-output features, potentially harming the distillation process.

Logit Standardization in Knowledge Distillation

- Shared Temperatures have its limitations:
  - forces the student to match the teacher's logits precisely, hindering the student's ability to learn within its capacity constraints
  - lead to misleading evaluations of student performance
- Z-score Logit Standardization applies to both teacher and student logits before the softmax function
  - Allows the student to learn and preserve the relative relationships between logits without requiring a strict magnitude match

#TrainingStrategy