Knowledge distillation in deep neural network

Presenter: Seunghyun Lee

인하대학교
INHA UNIVERSITY
1954

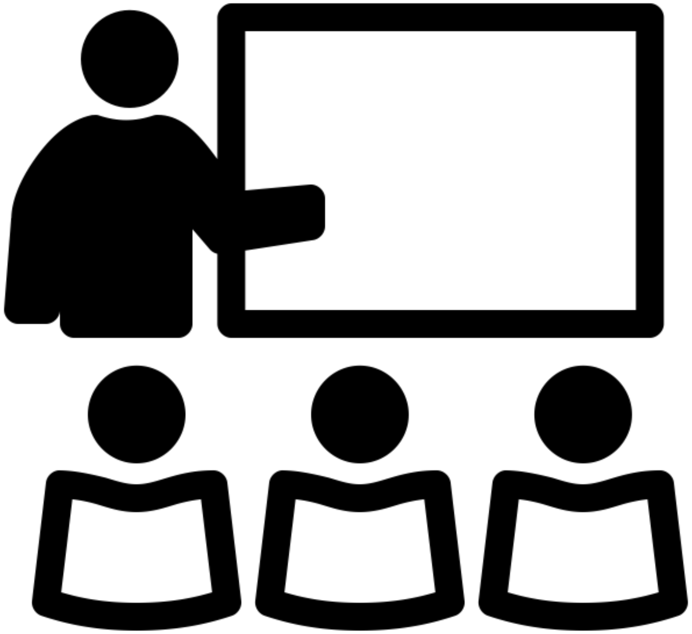
Computer Vision
& Image Processing
Laboratory

Contents

- Concept of Knowledge Distillation
- Response-based knowledge
    - Soft-logits – arxiv 2015
    - Selective Knowledge Distillation – TIP 2019
- Multi-connected network knowledge
    - FitNet – ICLR 2015
    - Attention Transfer – ICLR 2017
    - Activation Boundary – AAAI 2019
- Shared-representation Knowledge
    - FSP – CVPR 2017
    - KD-SVD – ECCV 2018
    - RKD – CVPR 2019
- Summary and Conclusion

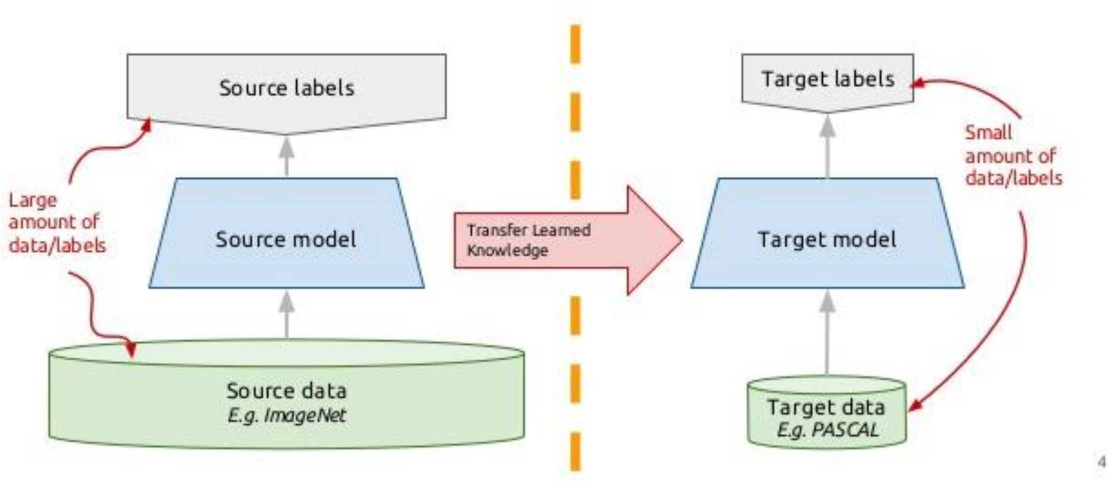
Concept of Knowledge Distillation

- Knowledge distillation
  - Distill a knowledge of large and complex network which called the teacher network.
  - Transfer it to a small and simple network which called the student network.



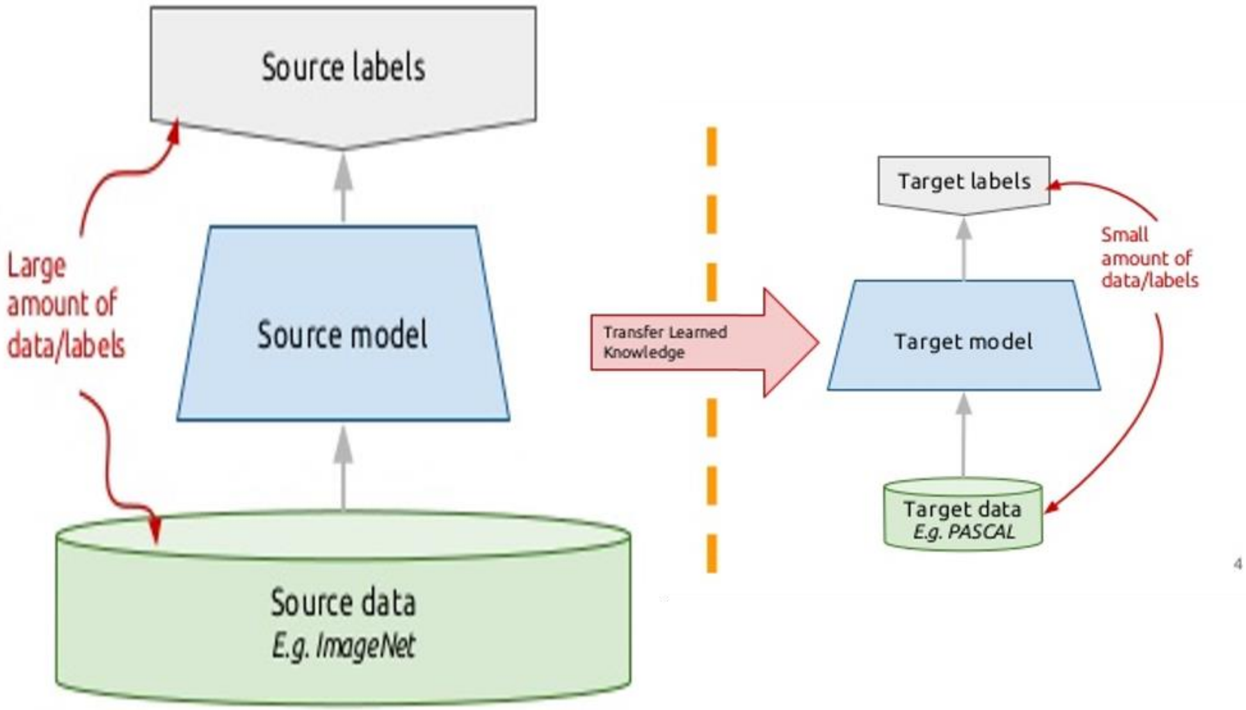
Concept of Knowledge Distillation

- Transfer learning
  - Train network by data from the source domain.
  - Finetune the network by data from the target domain.
    → Network's performance enhanced due to source domain data's information.
    → Usually, use a large dataset than the target.



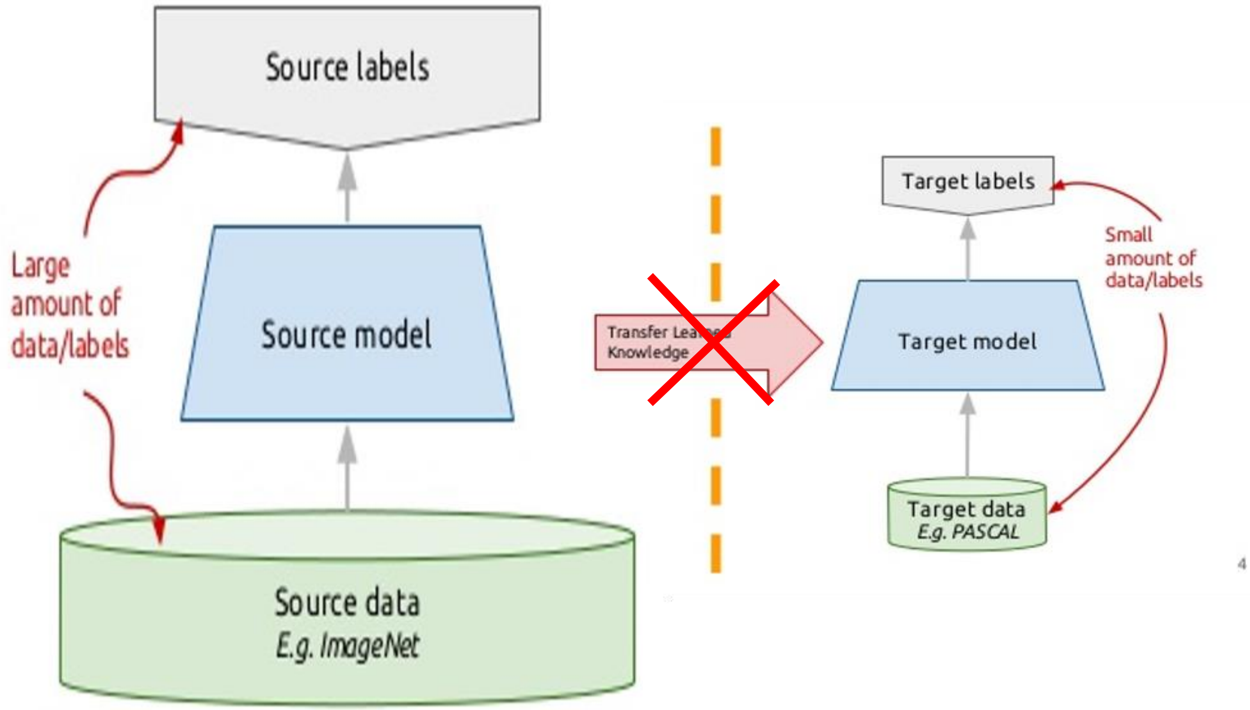
Concept of Knowledge Distillation

Is it better if use not only a large dataset but also a larger network?



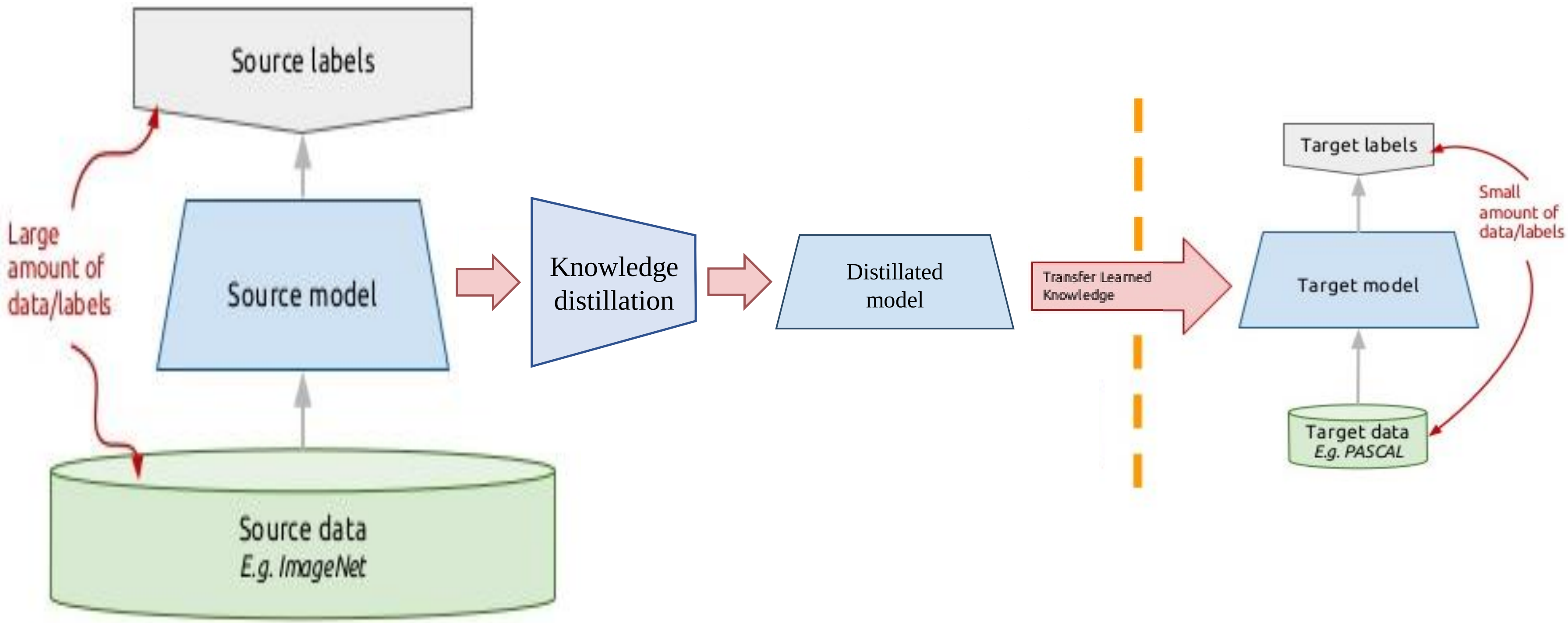
Concept of Knowledge Distillation

Is it better if use not only a large dataset but also a larger network?
→ Because the source and target network are different, information cannot be transferred.



Concept of Knowledge Distillation

- Knowledge distillation
  - Extract knowledge in a large network to make possible to transfer to a smaller network.



Concept of Knowledge Distillation

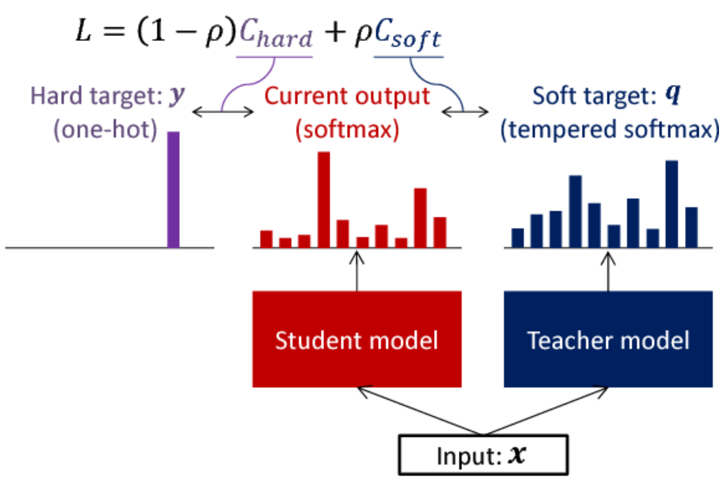
- Transfer learning
  → Method for transferring information to a target network from a source network.

- Knowledge distillation
  → Method for distillation to make teacher's information transferable.

** Key-point is defining the knowledge **

Soft-logits [1]

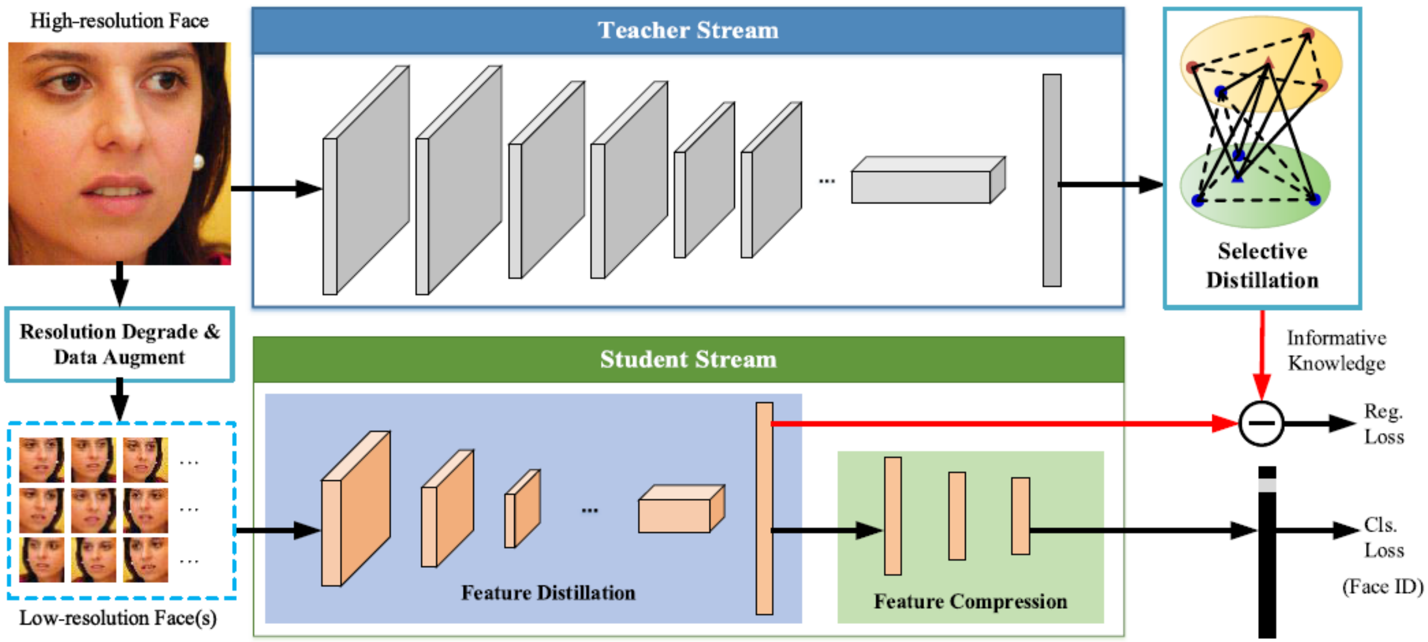
- Abstract
  - The paper which proposes knowledge distillation first.
  - Define knowledge as a soften teacher network's output.
- Pros
  - Because It is very easy to implement and handle, has been applied to various other methods.
    → Semi-supervised learning, combine with other KD method
- Cons
  - Knowledge's quality and quantity are too low.



[1] Geoffrey Hinton, et al. Distilling the knowledge in a neural network. arXiv:1503

Selective Knowledge Distillation [2]

- Abstract
  - The method for low-resolution face recognition using knowledge distillation.
  - Define teacher knowledge as embedded high-resolution face to make student network embed augmented low-resolution face well.
  - To give only useful information, select data which well embedded transfer.



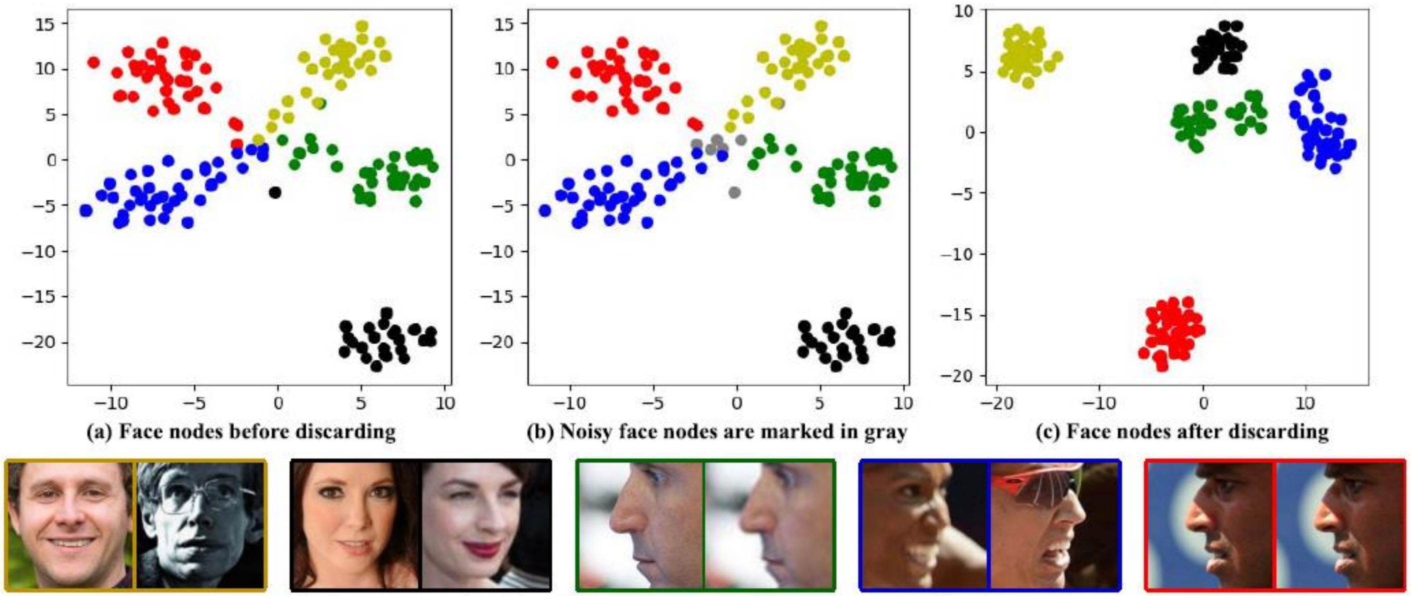
[2] Ge, Shiming, et al. "Low-resolution face recognition in the wild via selective knowledge distillation." IEEE TIP 2019.

Computer Vision & Image Processing Laboratory

Selective Knowledge Distillation

- Distillation mechanism
  1. Train student network by low-resolution face image.
  2. embedding dataset by teacher network and remove data not well embedded by graph cut.
  3. Fine-tune the student network by knowledge distillation with the selected dataset.

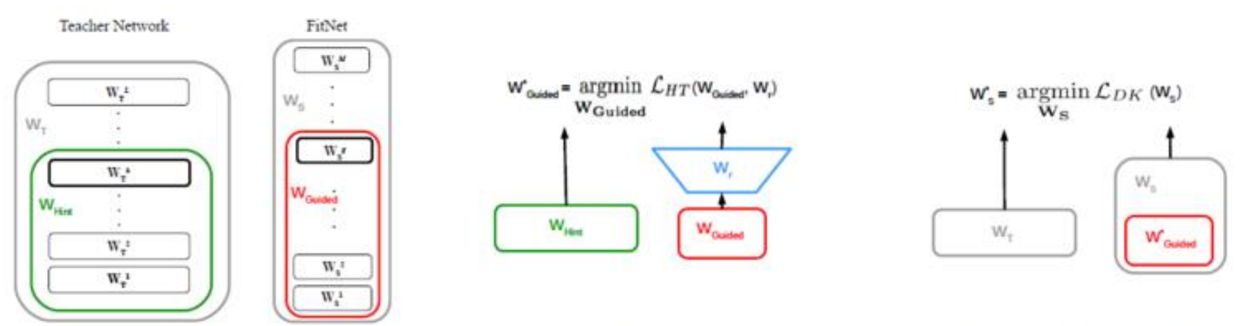
     → Quality of knowledge is improved cause of data's size and selection.



(d) The corresponding discarded face images (marked in the same color with face nodes)

FitNet [3]

- Abstract
  - Sensing multiple points that contain similar context information in teacher and student networks.
  - define loss function as L_2-distance of each feature map and initialize student
  - Finetune initialized student network.
- Pros
  - Quantity of knowledge is increased.
  - Gradients are well-propagated because of multiple connections.
- Cons
  - Student may be over-constraint by teacher's too sharp and complex knowledge.



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

[3] Adriana Romero, et al. Fitnets: Hints for thin deep nets. ICLR 2015.

Attention transfer [4]

- Abstract
  - Suppose spatial information is more important than feature information.
  - Inspired by that if some networks have high accuracy they have similar attention maps.

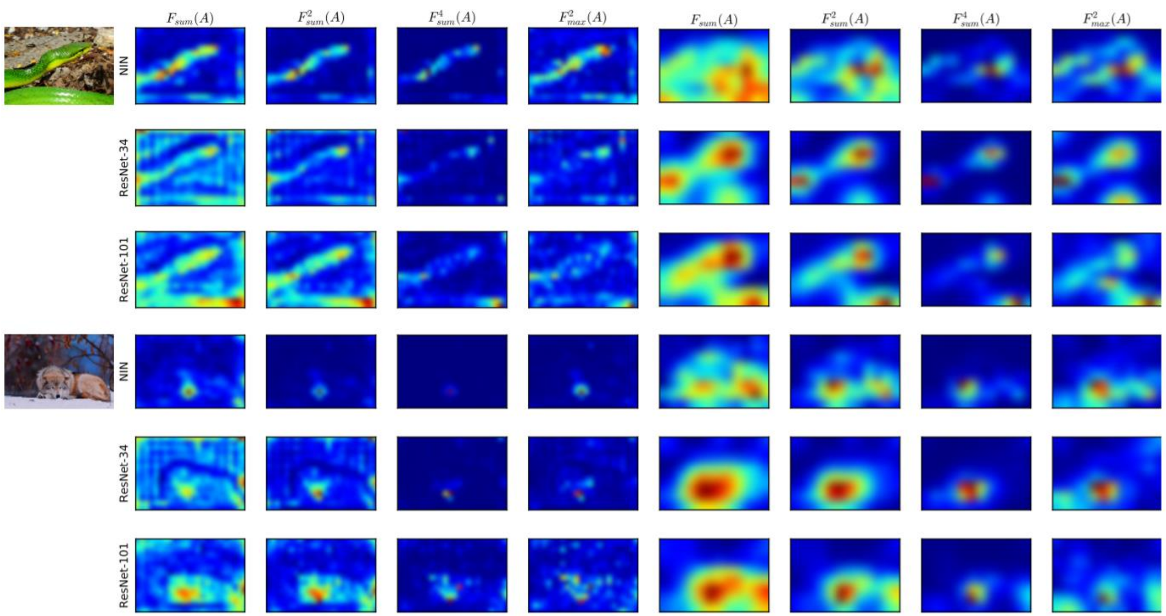
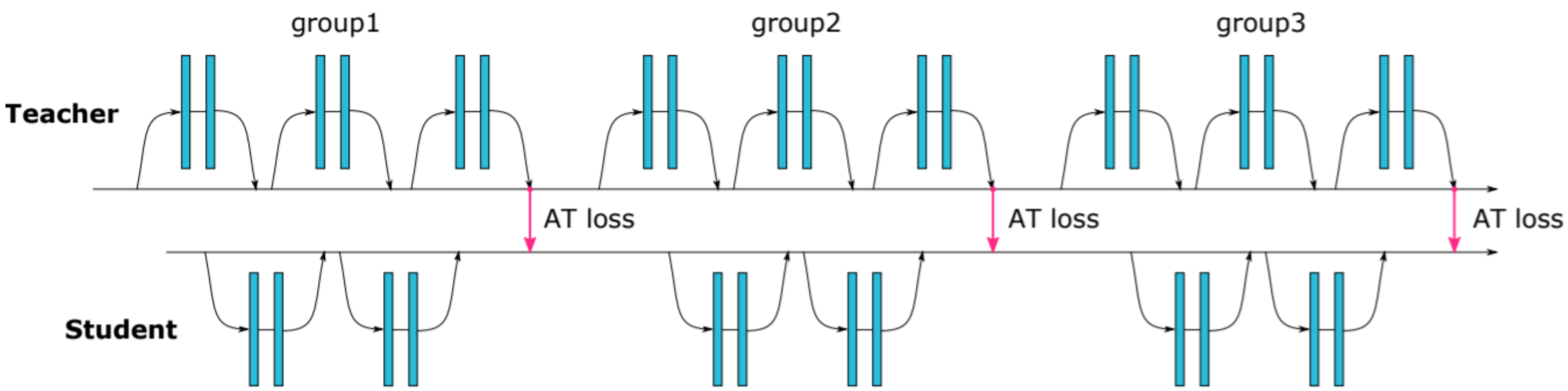


Figure 4: Activation attention maps for various ImageNet networks: Network-In-Network (62% top-1 val accuracy), ResNet-34 (73% top-1 val accuracy), ResNet-101 (77.3% top-1 val accuracy). Left part: mid-level activations, right part: top-level pre-softmax acivations

[4] Zagoruyko, Sergey et. al. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. ICLR 2017.

Attention transfer

- Distillation mechanism
  - Very similar to FitNet's training mechanism.
  - The only difference is defined knowledge, which is attention map computed by L_2-norm of each feature point.
    → Attention map is a 2-dimensional matrix, so teacher knowledge can be transferred irrespectively feature depth.
    → Attention map is much smoother than the original feature map, so the over-constraint problem is reduced.



Activation boundary [5]

- Abstract
  - The authors point out a problem of metric to compare each feature map in FitNet.
  - The authors suppose that classifier is a combination of the decision boundaries.
  - By L_2-distance it is hard to train decision boundary, so the other metric is needed.

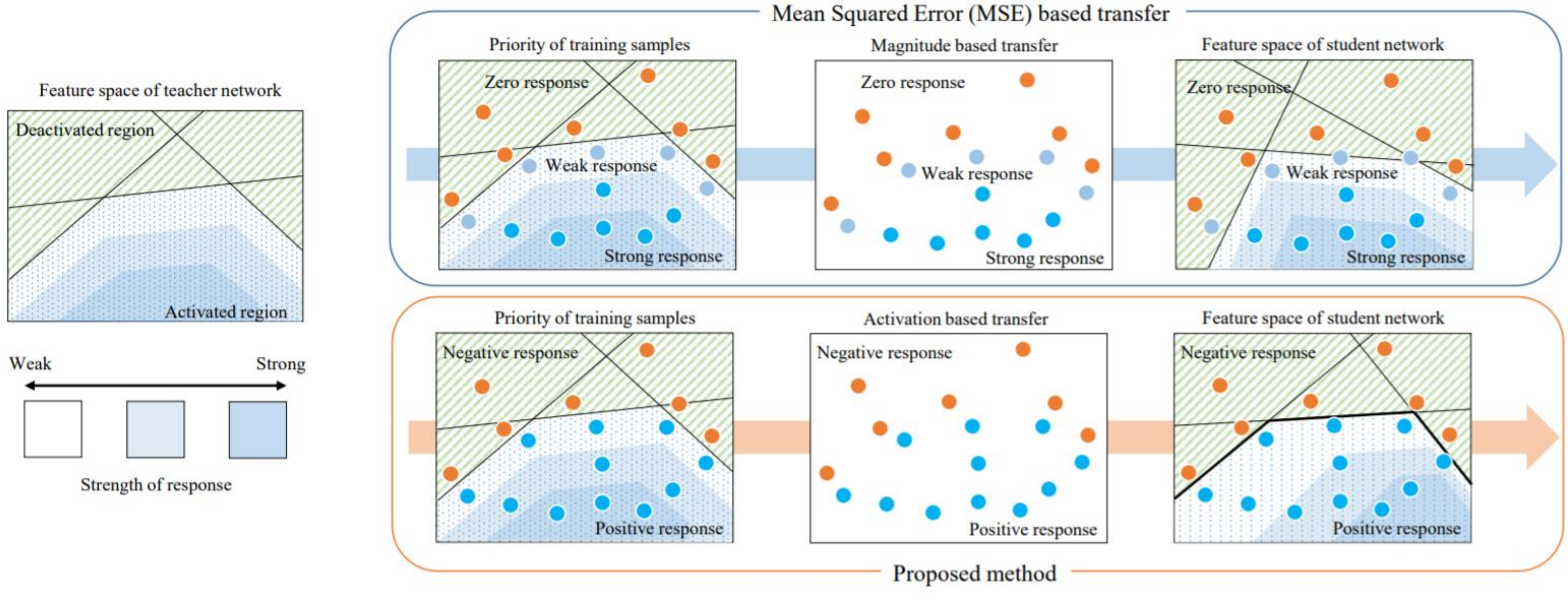


Figure 1: The concept of the proposed knowledge transfer method. The proposed method concentrates on the activation of neurons, not the magnitude of neuron responses. This concentration enables more precise transfer of the activation boundaries.

[5] Byeongho Heo, et. al. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. AAAI 2019

Activation boundary

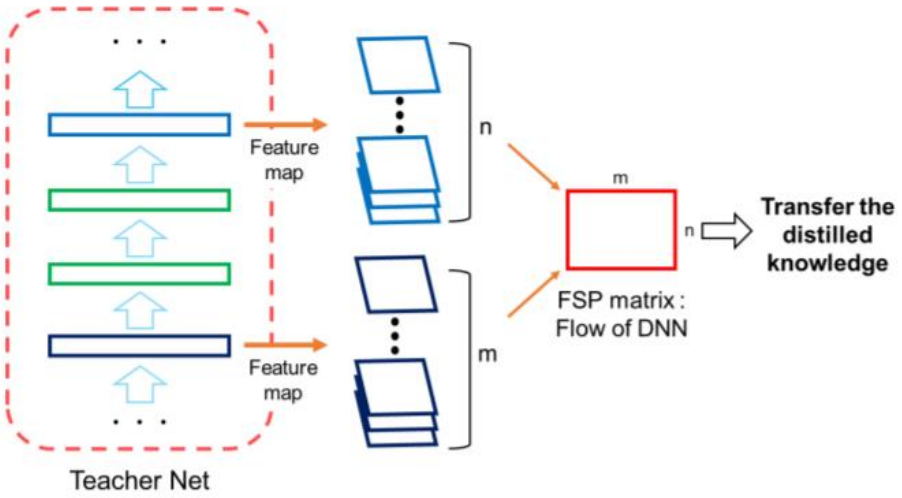
- Distillation mechanism
  - Very similar to FitNet's training mechanism.
  - Replace L_2-distance with Hinge loss which usually uses for SVM.
  - Define derivation function to make possible to train by back-propagation.

$$\mathcal{L}(\boldsymbol{I}) = \|\rho(\mathcal{T}(\boldsymbol{I})) \odot \sigma(\mu\mathbf{1} - \mathcal{S}(\boldsymbol{I})) + (\mathbf{1} - \rho(\mathcal{T}(\boldsymbol{I}))) \odot \sigma(\mu\mathbf{1} + \mathcal{S}(\boldsymbol{I}))\|_2^2$$

$$-\frac{\partial \mathcal{L}(\boldsymbol{I})}{\partial s_i} = \begin{cases} 2(s_i - \mu), & \text{if } \rho(t_i) = 1 \text{ and } s_i < \mu \\ -2(s_i + \mu), & \text{if } \rho(t_i) = 0 \text{ and } s_i > -\mu \\ 0, & \text{otherwise.} \end{cases}$$

Flow of solving procedure [6]

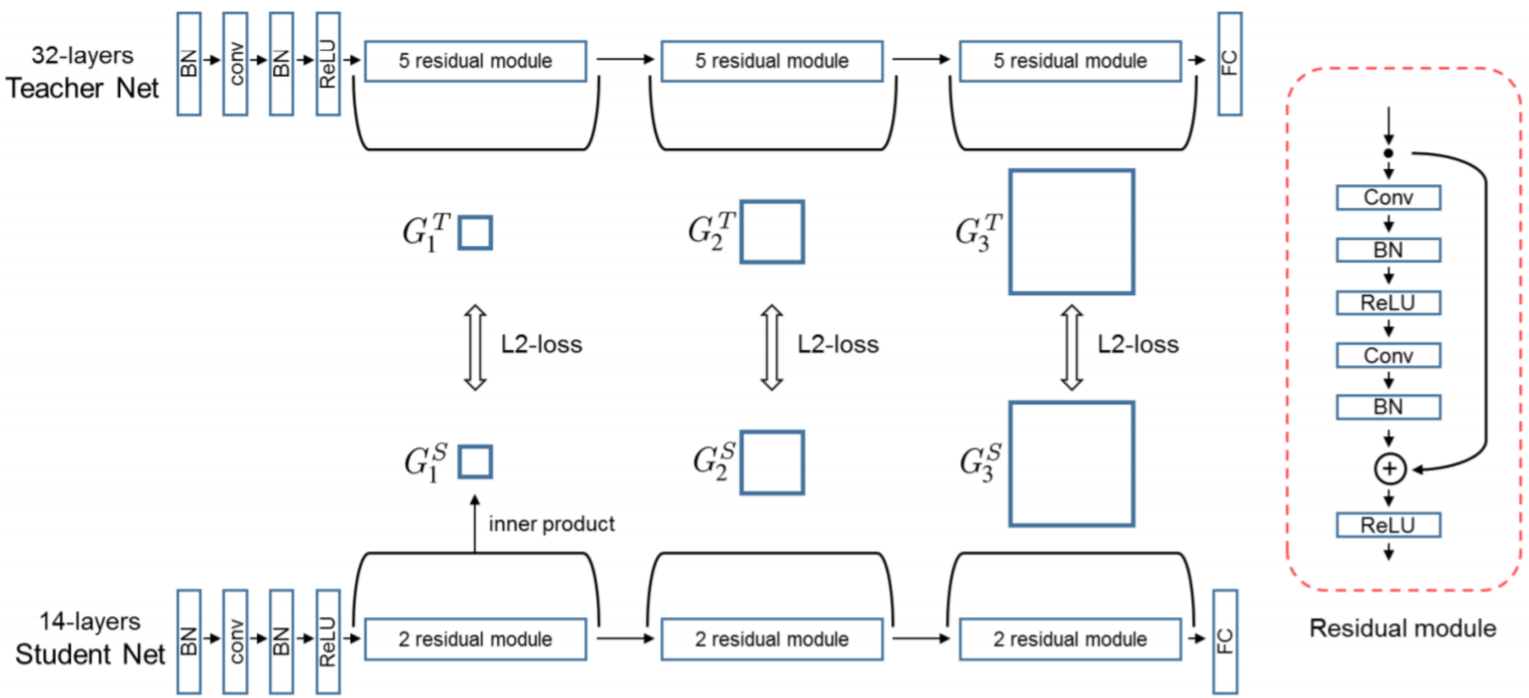
- Abstract
  - To solve FitNet's over-constraint problem, the authors define knowledge as shared-representation.
  - When sensing two points of a network and computing relation of them, the relation has information about feature transform which is the flow of solving procedure.



[6] Junho Yim, et. al. A gift from knowledge distillation: Fast optimization, network minimization, and transfer learning. CVPR 2017.

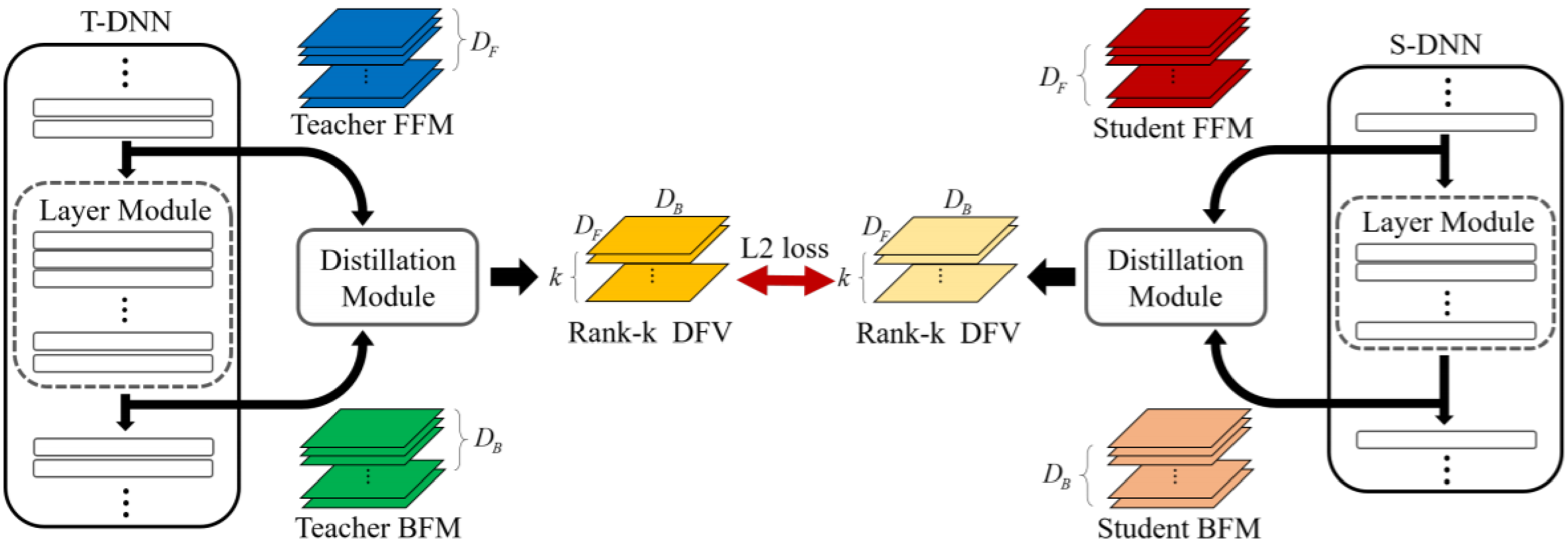
Flow of Solving Procedure

- Distillation mechanism
  - Sensing two points of each network.
  - Compute Grammian matrix which is a relation of two feature maps.
  - Initialize student network by minimizing the difference of knowledge.



Knowledge Distillation using SVD [7]

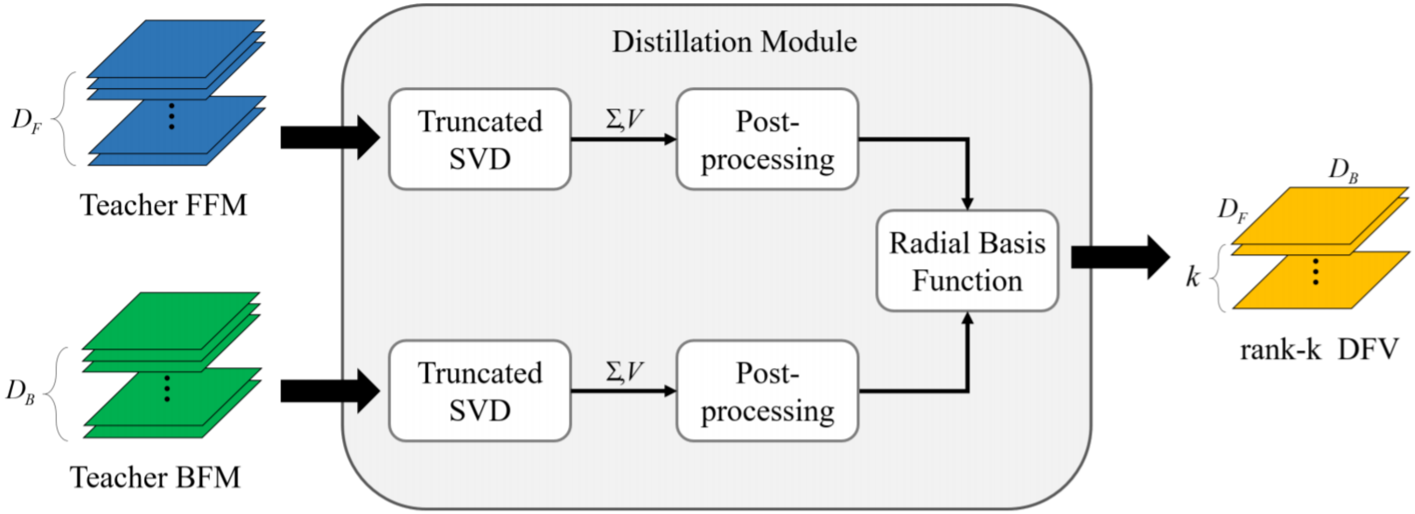
- Abstract
  - To get more important information and relation of feature maps, the authors use singular value decomposition and radial basis function.
  - Propose the adaptive constraint multi-task learning method to prevent over-constraint problem



[7] Seung Hyun Lee, et. al. Self-supervised knowledge distillation using singular value decomposition. ECCV 2018

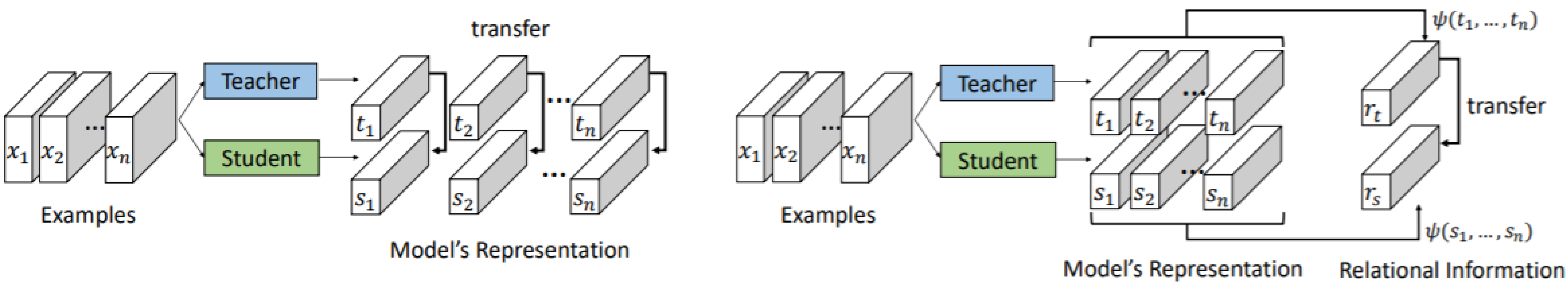
Knowledge Distillation using SVD

- Distillation mechanism
  - Sensing two points of each network.
  - Compress each feature map by SVD, post-process for removing bad property of singular vectors and compute relation of singular vectors by RBF.
  - Clip gradients of transfer learning by norm of gradients of main-task learning.



Relational Knowledge Distillation [8]

- Abstract
  - Point out all of distillation methods cannot distill information about inter-data relation.
  - If student network gets information inter-data relation, student network can embed dataset like teacher network.

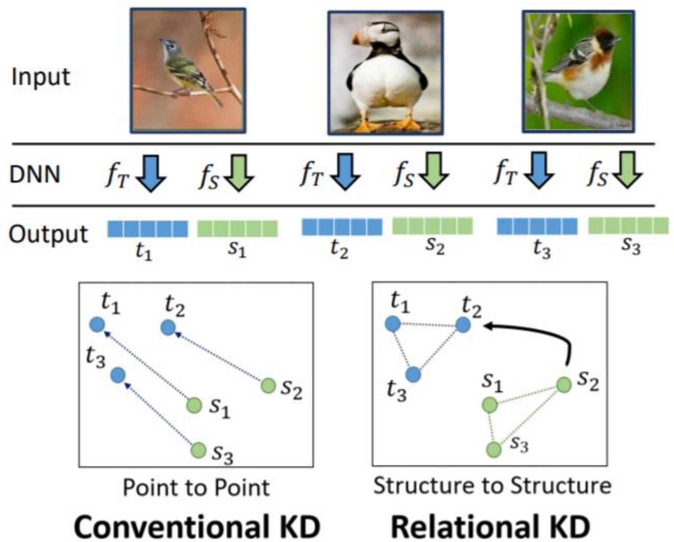


Individual Knowledge Distillation **Relational Knowledge Distillation**

[8] Wonpyo Park, et. al. Relational Knowledge Distillation. CVPR2019

Relational Knowledge Distillation

- Distillation mechanism
  - Sense embedded feature vector such as feature extractor or network's output
  - Compute distance-wise relation and angle-wise relation of each feature vector.
  - Train student network by multi-task learning.



$$\psi_{\mathrm{D}}(t_i, t_j) = \frac{1}{\mu} \|t_i - t_j\|_2$$

$$\psi_{\mathrm{A}}(t_i, t_j, t_k) = \cos \angle t_i t_j t_k = \langle \mathbf{e}^{ij}, \mathbf{e}^{kj} \rangle$$

Summary and Conclusion

- Response-based knowledge
  - Very simple so knowledge's quality and quantity are too low but easy to handle and modify.
  - SOTA methods focus to extract better knowledge.
  - Soft-logits, Selective knowledge etc.
- Multi-connected network knowledge
  - Distill a large amount of knowledge, but it may cause over-constraint.
  - SOTA methods focus to soften teacher's knowledge.
  - FitNet, Attention transfer, Activation boundary etc.
- Shared-representation knowledge
  - Distill a large amount of softened knowledge, but computational cost is much larger than others.
  - FSP, KD-SVD, RKD etc.

Summary and Conclusion

- Nothing is completely superior. So we have to choose a proper one.