Graph-based Knowledge Distillation by Multi-head Attention Network

*Seunghyun Lee
lsh910703@gmail.com

Byung Cheol Song
bcsong@inha.ac.kr

Inha University, Incheon, Republic of Korea

Contents

- Background
- Problem Statement
- Multi-head Graph Distillation
- Experimental Results
- Conclusion

Background (1/5)

- Light-weighting of network
  - CNN is useful for many tasks, but its cost of computing and memory is still massive.
  - A lot of techniques for light-weighting CNNs have been proposed.
    ex) Pruning, quantization, **knowledge distillation,** etc.
- Knowledge distillation (KD)
  - **Achieve optimal performance** from a small student network (SN) by distilling the knowledge of a large teacher network (TN) and transferring the distilled knowledge to the small SN.
  - Distilled knowledge can be applied for other purposes such as semi-supervised learning and transfer learning.

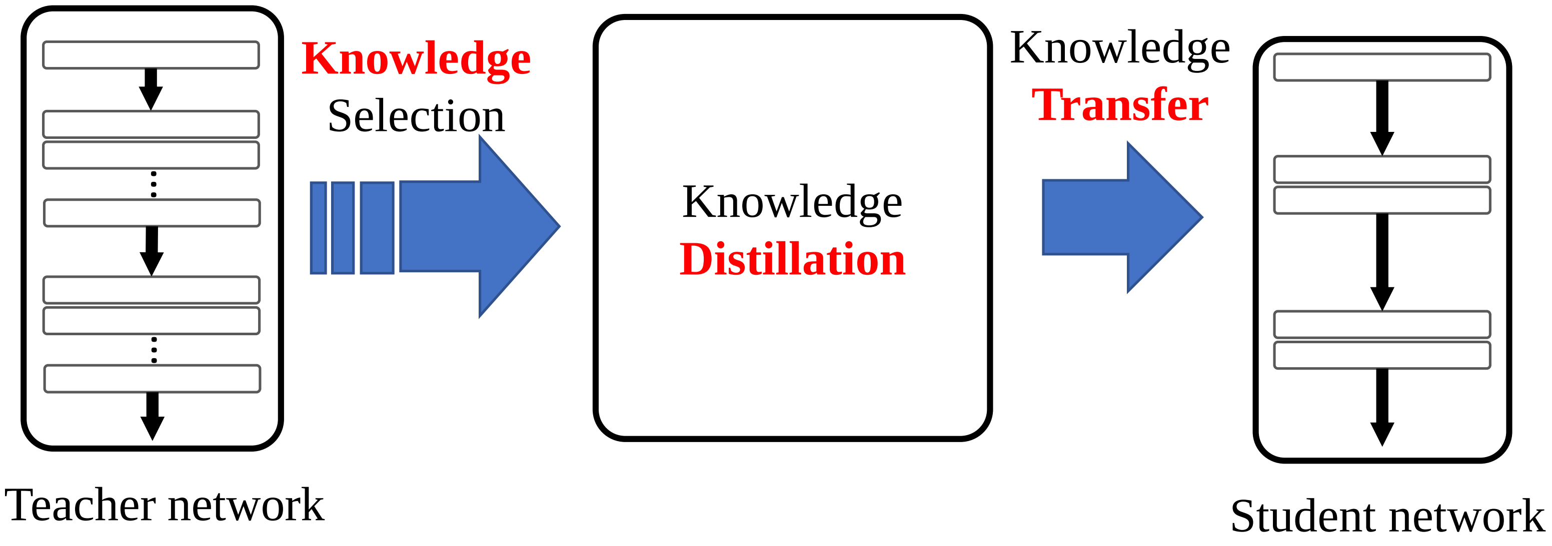
Background (2/5)

- Knowledge distillation procedure
  - Consists of three important components.

    → **Selecting TN's knowledge** to distill,

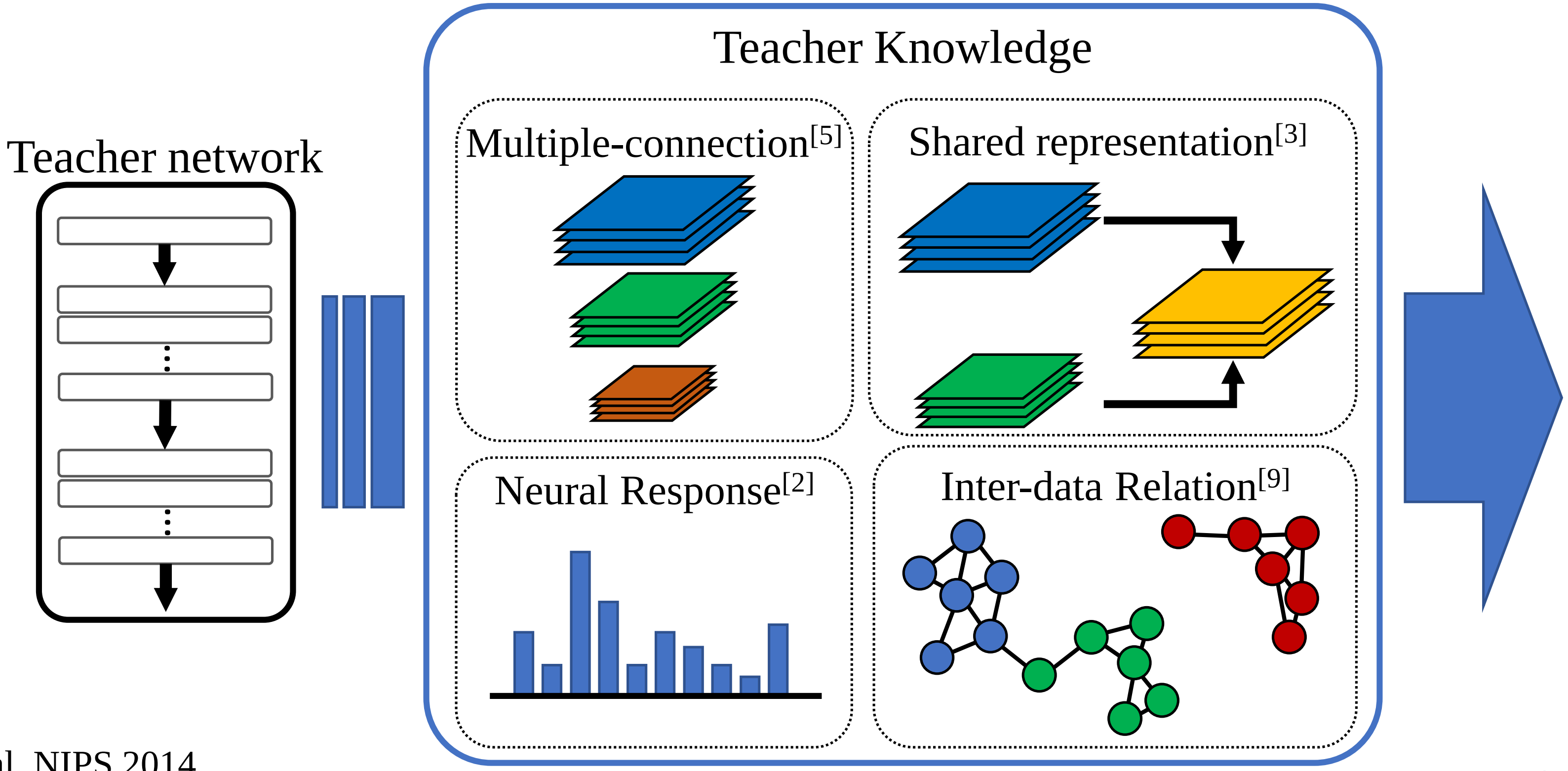
    → **Distilling TN's knowledge**,

    → **Transferring knowledge** to SN.



Background (3/5)

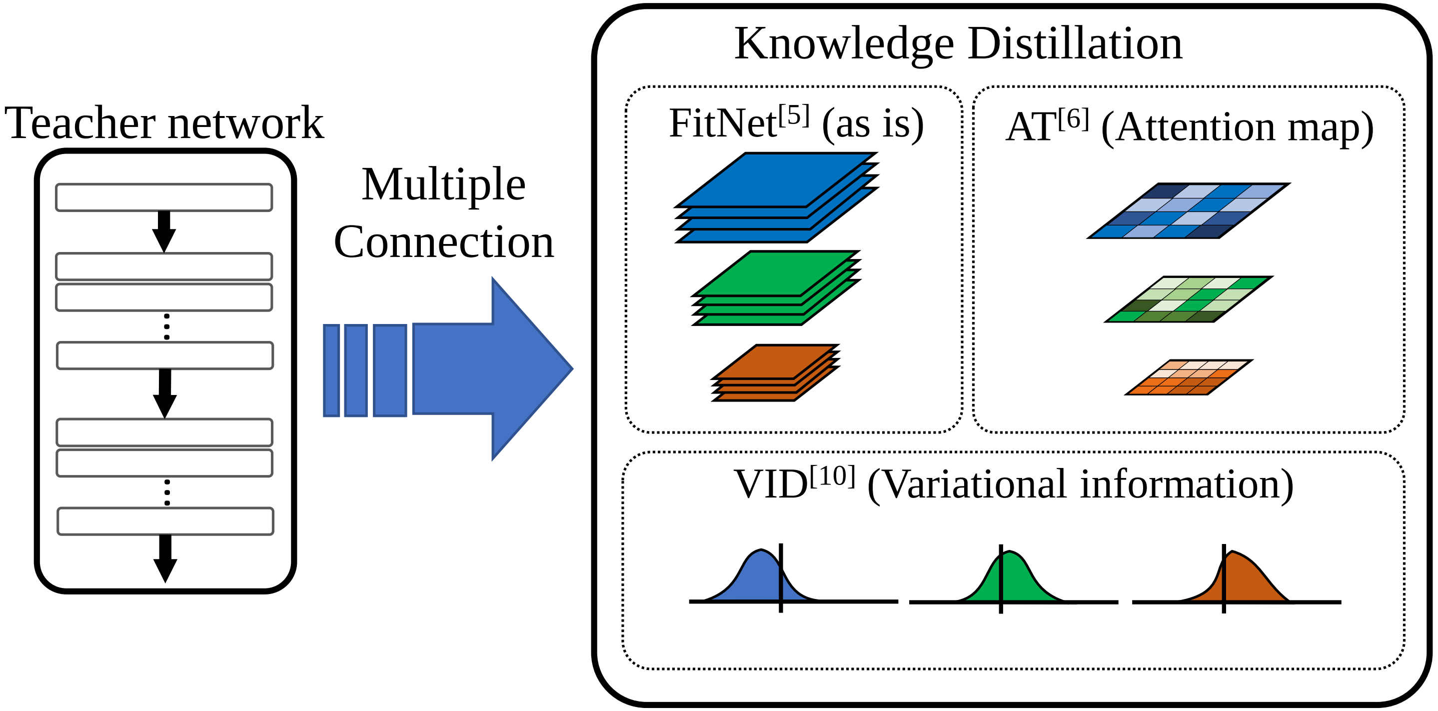
- Selecting TN's knowledge to distill
  - Extract the TN's feature, or just determine the way for distillation.



[2] Hinton et al. NIPS 2014
[3] Yim et al. CVPR2017
[5] Romeo et al. ICLR2015
[9] Park et al. CVPR2019

Background (4/5)

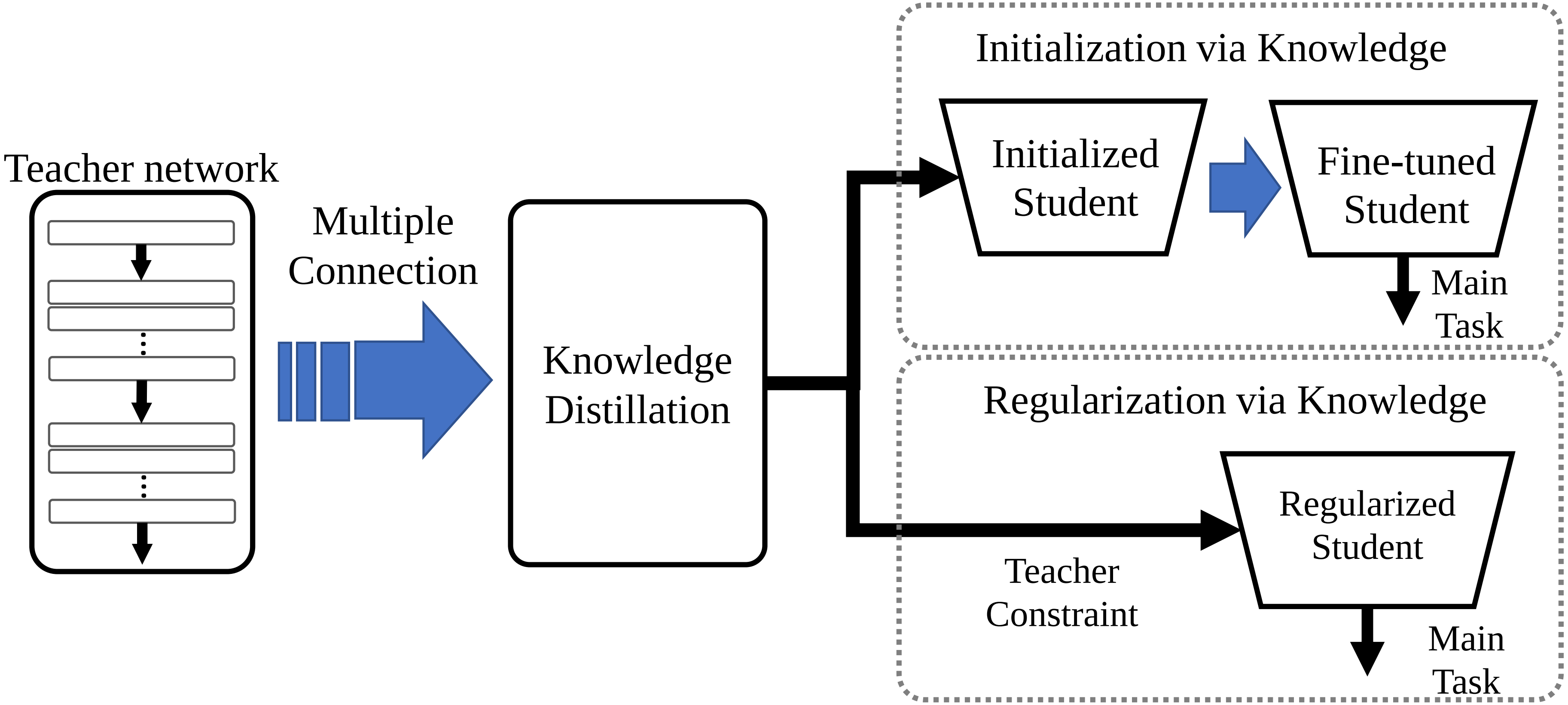
- Distilling TN's knowledge
  - Soften the teacher's knowledge, or Construct the feature which represents the selected knowledge.



[5] Romeo et al. ICLR2015
[6] Sergey et al. ICLR2017
[10] Ahn et al. CVPR2019

Background (5/5)

- Transferring the distilled knowledge to SN
  - Initialize or regularize the SN using the TN's knowledge.



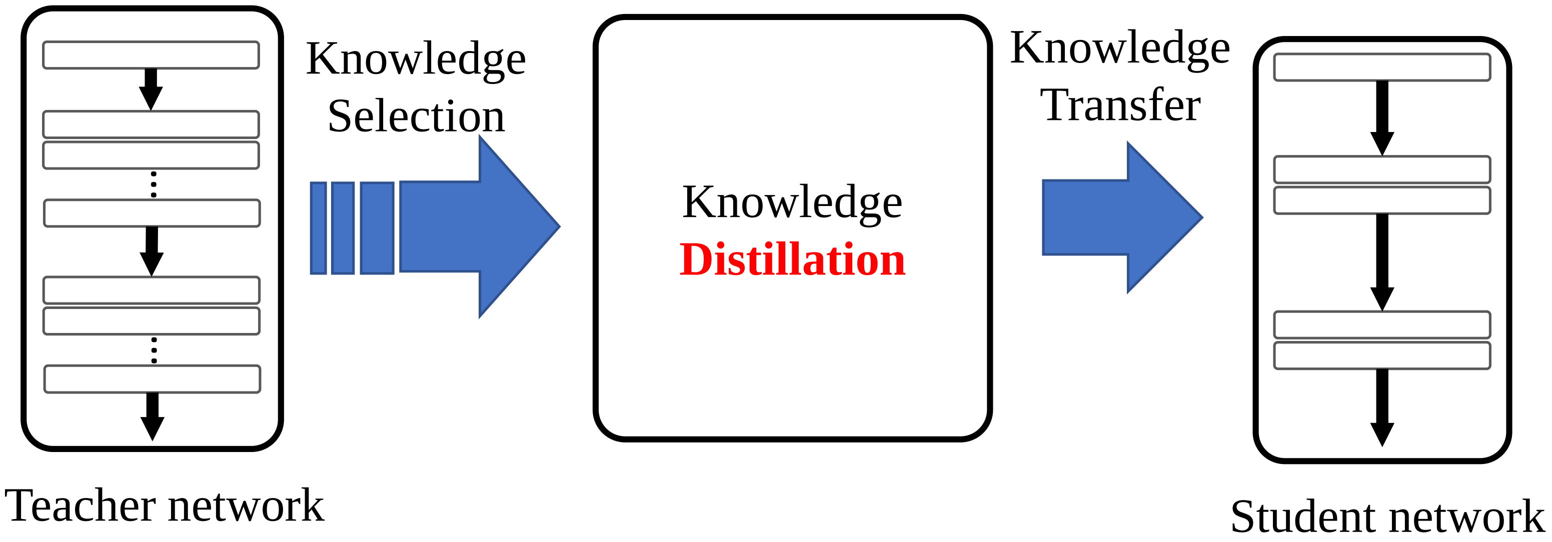
Problem Statement (1/4)

- Knowledge distillation procedure
  - Consists of three important components.

    → **Selecting TN's knowledge** to distill,

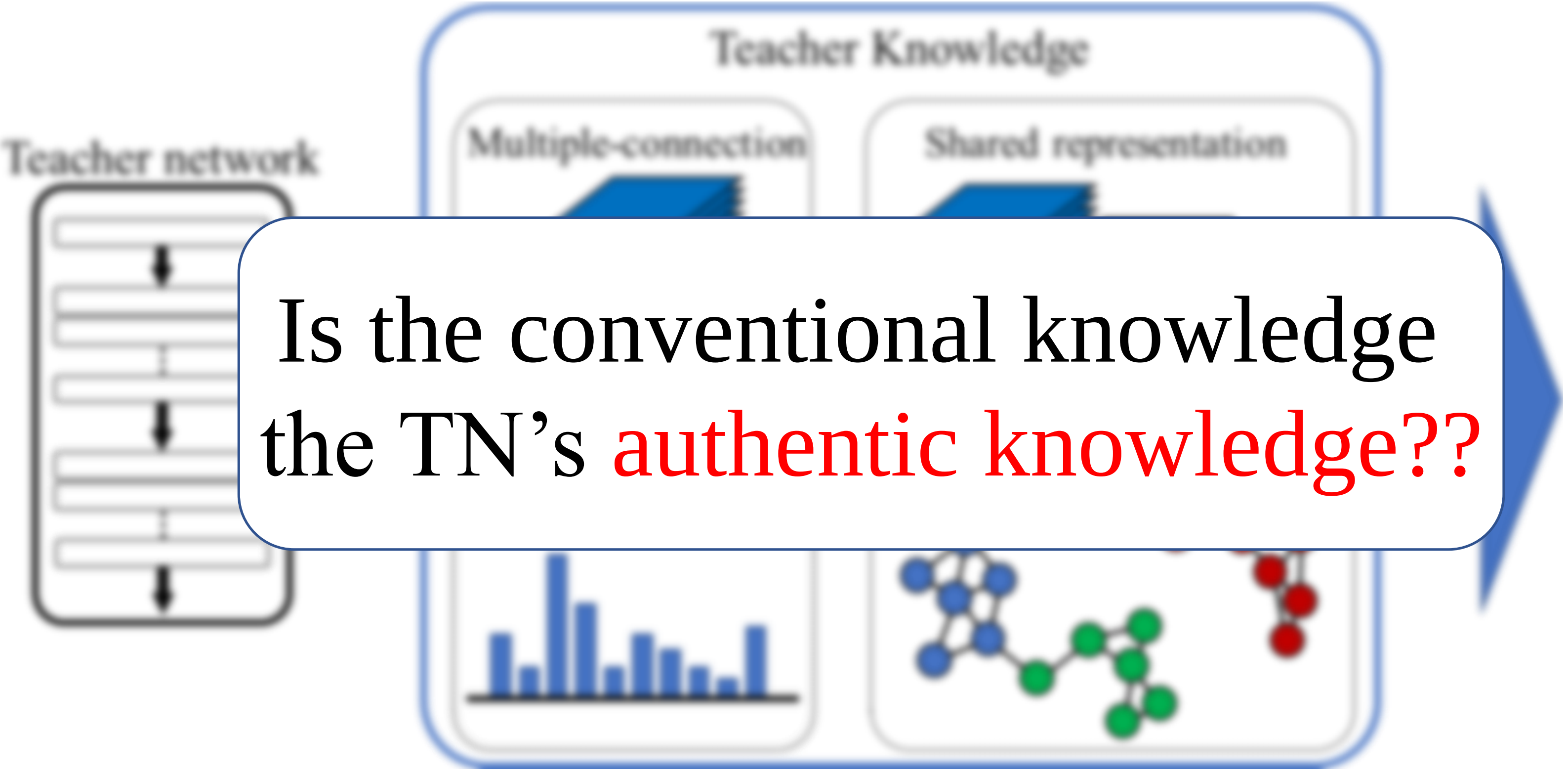
    → **Distilling TN's knowledge**,

    → **Transferring knowledge** to SN.



Problem Statement (2/4)

- Selecting TN's knowledge to distill
  - Extract the TN's feature, or just determine the way for distillation.



Is the conventional knowledge
the TN's authentic knowledge??

Problem Statement (3/4)

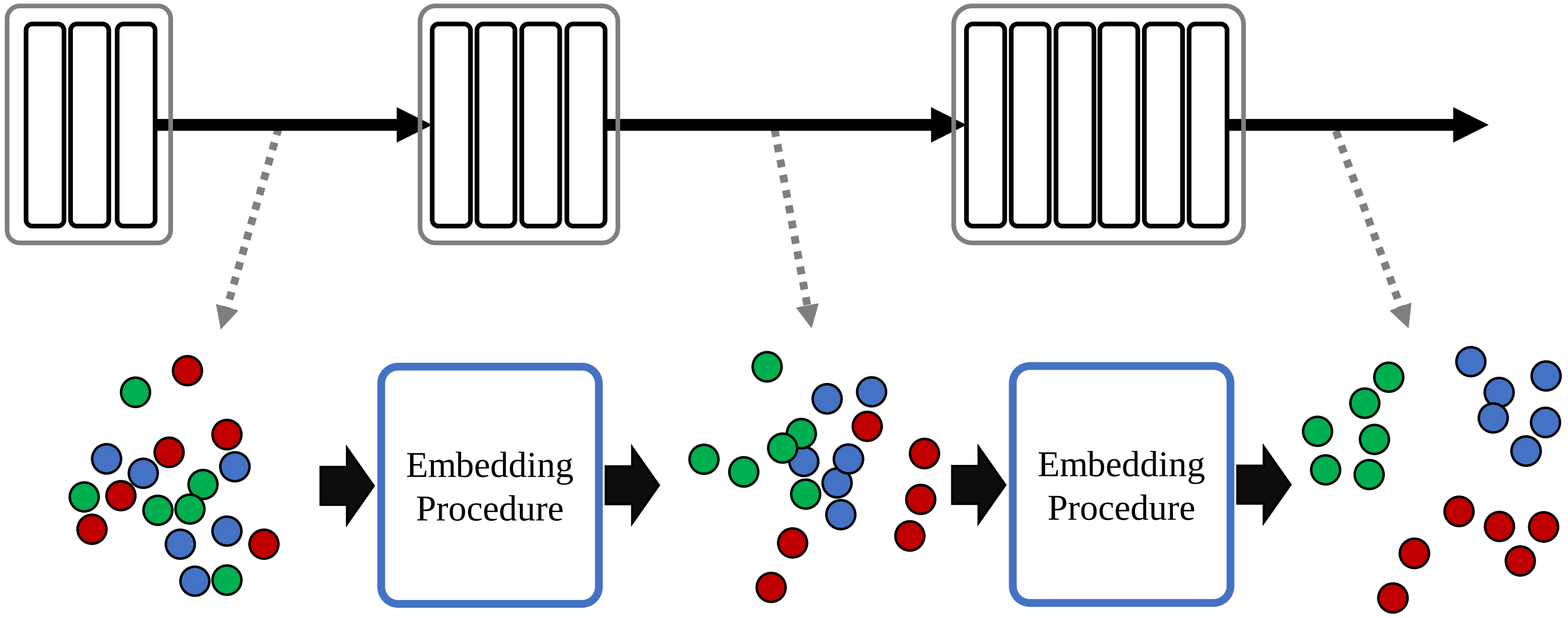
- Limitations of the previous approaches
  - Most of the previous methods focus on **How** to distill knowledge, not **What** to distill.
  - All type of knowledge is not still acceptable as a neural network's knowledge.
    - Neural response & Multi-connection : Too naive
    - Shared-representation : Cannot find inter-data relation
    - Inter-data relation knowledge : Only focus on the last embedded space

- <u>**Problem definition**</u>
  - Find **the <u>knowledge</u> which coincides with <u>neural network's purpose</u>**.

Problem Statement (4/4)

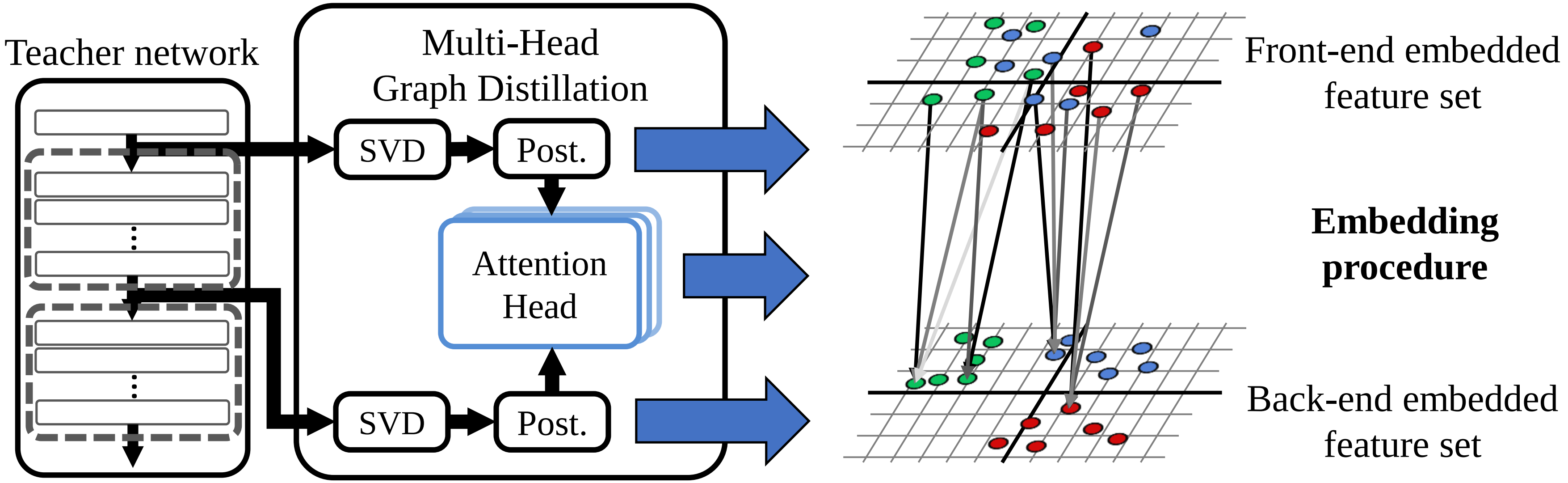
- Neural network's purpose
  - **Embed high-dimensional data to low-dimension** for easier analysis.
  - A good teacher teaches not only answer but **how to solve**.

    → **Embedding procedure** is the real knowledge of the neural network.



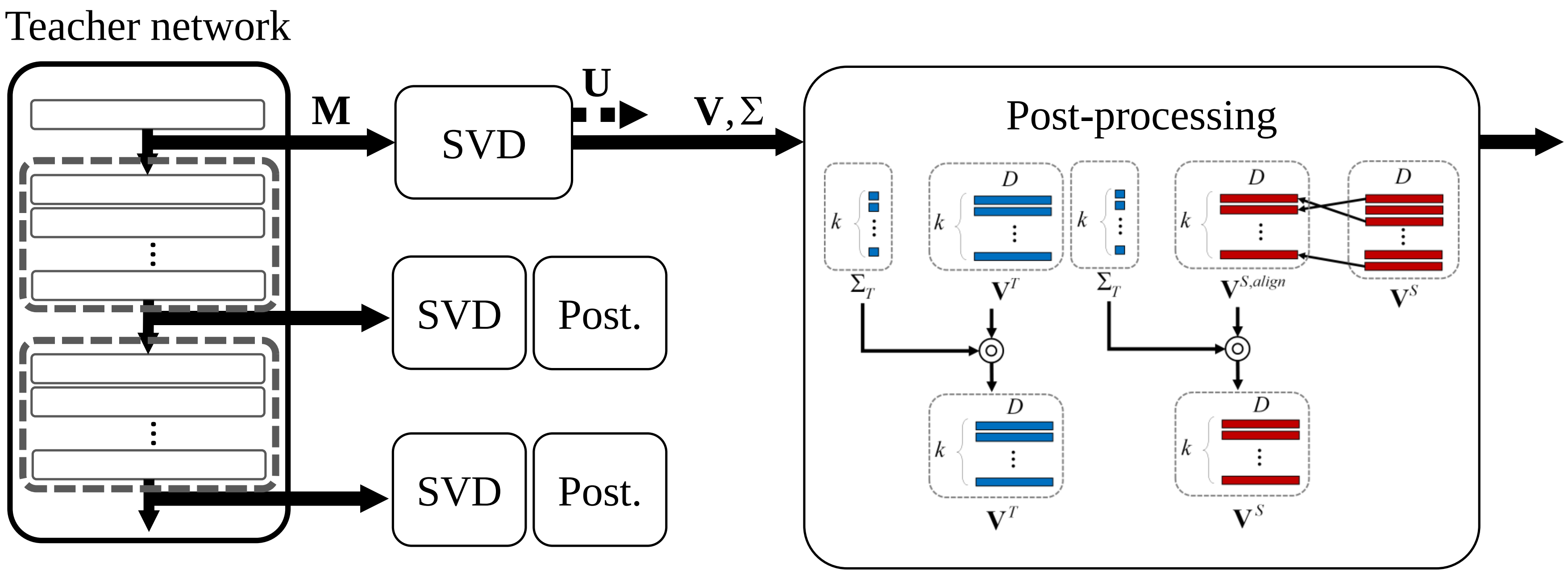
Multi-head Graph Distillation (1/5)

- Distill **embedding procedure knowledge,** i.e., the core information of neural network.
- Apply **SVD** and **attention network** to extract the feature map's relation, which is hard to apprehend.
- Transfer the knowledge via **multi-task learning** to supply TN's knowledge continuously.



Multi-head Graph Distillation (2/5)

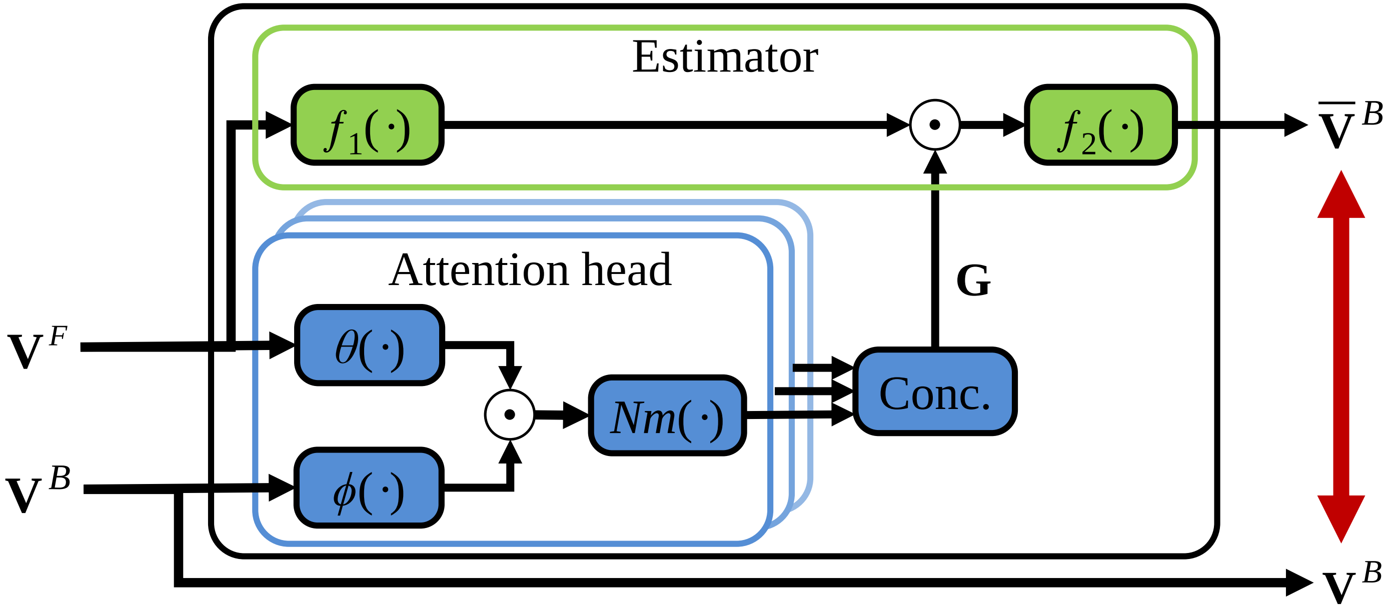
- Compressing feature maps by SVD
  - Feature map's dimension is too high to compute relation between them.
    → So, compress feature maps by SVD.
  - Apply post-processing to make them able to transfer. [1]



[1] Lee et al. ECCV 2018

Multi-head Graph Distillation (3/5)

- Multi-head attention network
  - **Estimator** which estimates back-end singular vector ($\mathbf{V}^B$) using front-end singular vector ($\mathbf{V}^F$).
  - **Attention head** which enhances the estimator's feature vector to make it easy to estimate $\mathbf{V}^B$.

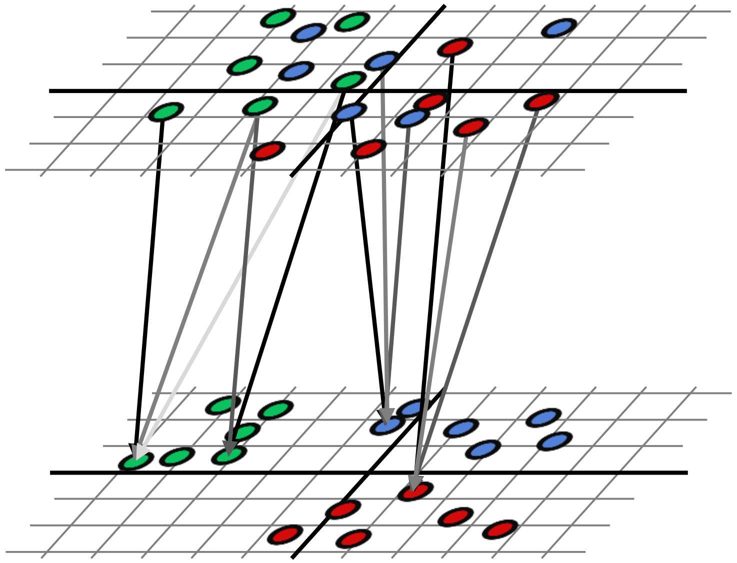


Multi-head Graph Distillation (4/5)

- Attention map as **Graph-based Knowledge**
  - Attention heads extract the relation between $\mathbf{V}^F$ and $\mathbf{V}^B$ to enhance $\mathbf{V}^F$ to estimate $\mathbf{V}^B$ easily.

    → Give more attention to the $\mathbf{V}^F$ which is embedded into similar points.

    → **Embedding procedure** is expressed by **graph-form**

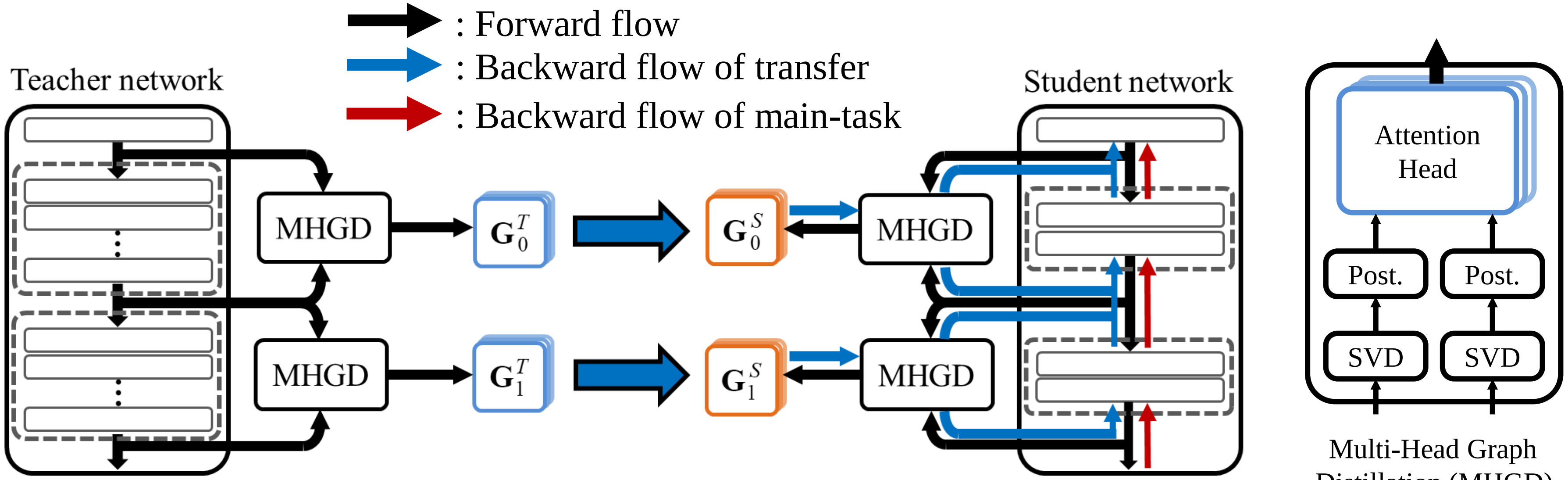


Graph-based Knowledge

Multi-head Graph Distillation (5/5)

- Transfer of graph-based knowledge
  - Adaptive constraint multi-task learning via gradient clipping [1].
  - Transfer the TN’s knowledge as much as possible without over-regularization.

$$\left(\frac{\partial\Theta}{\partial L_{KD}}\right)_{clipped} = \frac{1}{1+\exp(-\tau)}\frac{\partial\Theta}{\partial L_{transfer}}, \text{ where } \tau = \max\left(1, \left\|\frac{\partial\Theta}{\partial L_{main}}\right\|_2 \Big/ \left\|\frac{\partial\Theta}{\partial L_{transfer}}\right\|_2\right)$$



Multi-Head Graph Distillation (MHGD)

[1] Lee et al. ECCV 2018

Experimental Results (1/6)

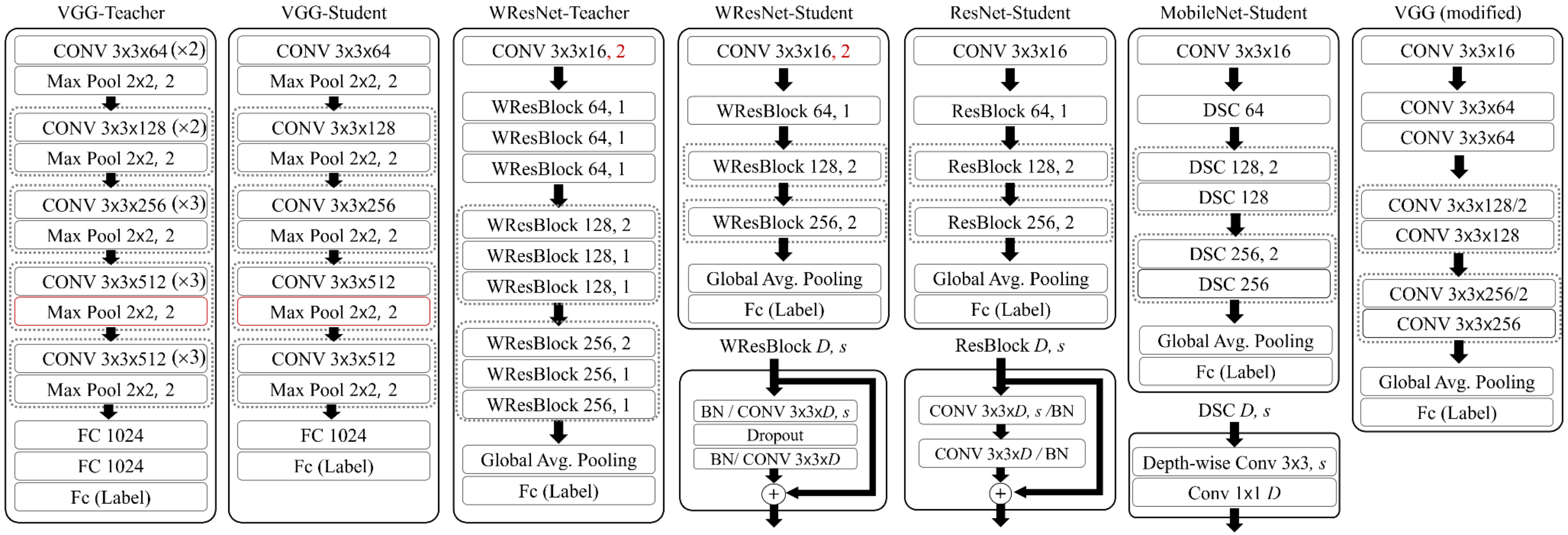
- Experiment setup
  - Network architectures
    - WResNet, VGG, ResNet, MobileNet
  - Datasets
    - CIFAR100, TinyImageNet
  - Previous methods

Method	Knowledge	Transfer method
Soft-logits [2]	Neural response	Multi-task learning
FSP [3]	Shared representation	Initialization
AB [4]	Multi-connection	Initialization
KD-SVD [1]	Shared representation	Multi-task learning
MHGD	Embedding procedure	Multi-task learning

[1] Lee et al. ECCV2018
[2] Hinton et al. NIPS 2014 Deep Learning Workshop
[3] Yim et al. CVPR2017
[4] Heo et al. AAAI2019

Experimental Results (2/6)

- Experiment setup
  - Sensing feature map from bold arrow of each architecture
  - For TinyImageNet, we added pooling layer that is marked redbox.



Experimental Results (3/6)

- Small network enhancement
  - KD-SVDF : Transfer singular vector **as is (multiple connection)**.
  - KD-SVD : Transfer singular vector's **shared representation**.
  - MHGD : Transfer singular vector's **embedding procedure**.

Network		FLOPS (M)	Params (M)
VGG	TN (VGG16)	143.7	11.83
	SN (VGG7)	17.6 (**12.2%**)	17.6 (**18.5%**)
WResNet	TN (WResNet22-4)	374.2	0.417
	SN (WResNet10-4)	93.2 (**24.9%**)	0.1404 (**33.7%**)

Performance comparison of several KD methods for **CIFAR100**.

Method	Teacher	Student	Soft-logits	FSP	AB	KD-SVD	KD-SVDF	MHGD
VGG	67.99	59.97	60.95	61.87	64.56	64.25	64.38	**67.02**
WResNet	77.22	71.62	71.88	71.57	72.23	71.83	71.82	**72.79**

Performance comparison of several KD methods in **TinyImageNet**.

Method	Teacher	Student	Soft-logits	FSP	AB	KD-SVD	KD-SVDF	MHGD
VGG	56.30	52.40	53.78	54.85	54.99	55.33	55.35	**56.35**
WResNet	61.31	55.91	56.00	56.04	56.53	55.72	55.95	**56.90**

Experimental Results (4/6)

- Knowledge transfer according to architecture
  - Even though transferring knowledge to SN that is different from TN's architecture, the proposed method outperforms the others.
  - In case of ResNet, most of distillation methods is failed to improve student network, but proposed method successes.

Performance comparison of various KD methods with **WResNet as the TN**.

Method	Student	Soft-logits	FSP	AB	KD-SVD	MHGD
VGG	69.76	70.51	69.44	71.24	70.31	**71.52**
MobileNet	66.18	67.35	60.35	67.84	67.03	**68.32**
ResNet	71.57	71.81	70.40	71.55	71.55	**72.74**

Experimental Results (5/6)

- Each attention heads extract different embedding information.
  - More attention heads tend to produce much knowledge.
  - However, too many attention heads may cause over-constraint.

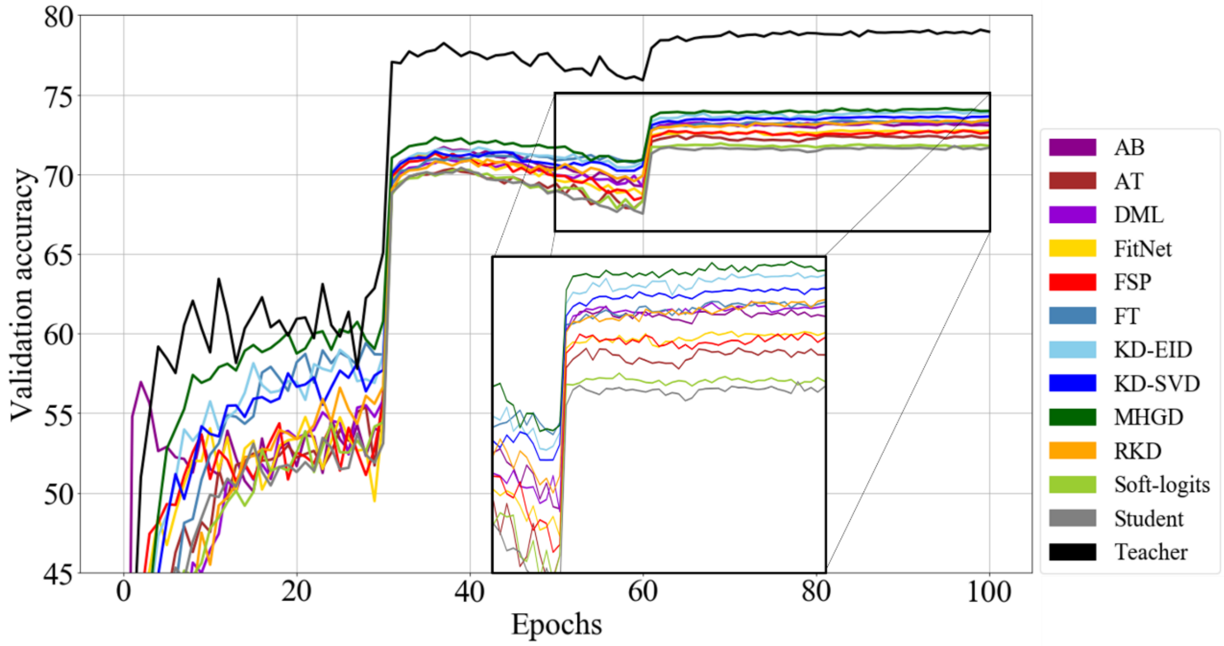
The performance change according to the number of attention heads.

num_head	0 (Student)	1	2	4	8	16
Accuray	59.97	65.71	66.41	67.01	**67.02**	66.70

Experimental Results (6/6)

- Comparison with the state-of-the art methods

Method	Student	Teacher	Soft-logits[2]	FitNet[5]	AT[6]	FSP[3]
Accuracy	71.76	78.96	71.79	72.74	72.31	72.65
Method	DML[7]	KD-SVD[1]	FT[8]	AB[4]	RKD[9]	MHGD
Accuracy	73.27	73.68	73.35	73.08	73.40	**73.98**



[1] Lee et al. ECCV2018
[2] Hinton et al. NIPS 2014
[3] Yim et al. CVPR2017
[4] Heo et al. AAAI2019
[5] Romeo et al. ICLR2015
[6] Sergey et al. ICLR2017
[7] Zhang et al. CVPR2018
[8] Kim et al. NeurIPS2018
[9] Park et al. CVPR2019

Conclusion

- Analyze previous knowledge distillation methods and point out their fundamental issue.
  →**No knowledge about embedding procedure** that is the purpose of neural networks **yet**.

- Propose a novel **algorithm to extract embedding procedure knowledge via attention networks**.

Thank you

Questions?

https://github.com/sseung0703/KD_methods_with_TF

Visit poster 141.