KNOWLEDGE TRANSFER VIA DISTILLATION OF ACTIVATION BOUNDARIES FORMED BY HIDDEN NEURONS

(Byeongho Heo, Minsik Lee, Sangdoo Yun, Jin Young Choi)

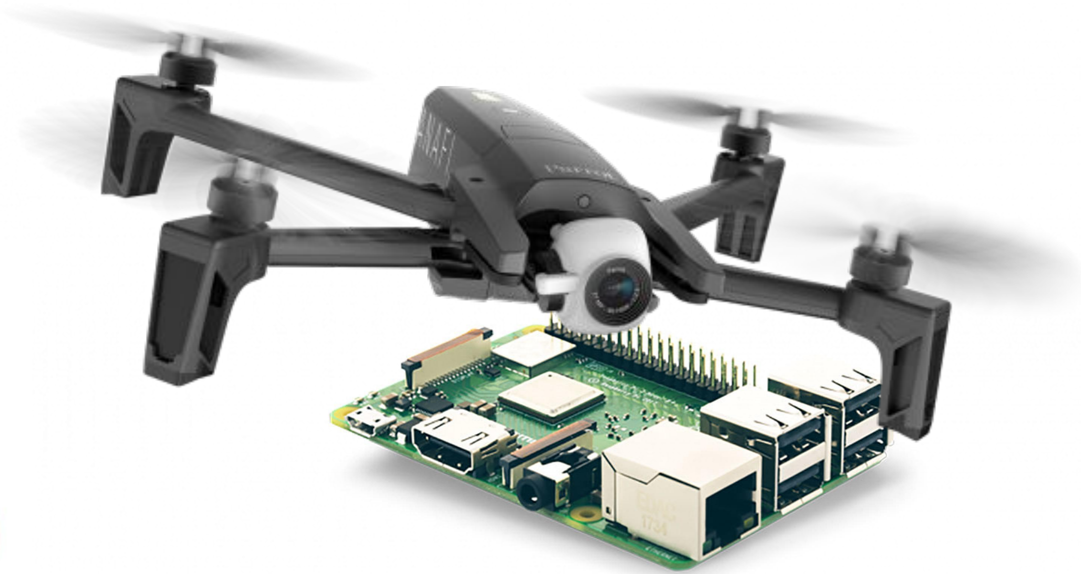
AIC OPT6 - Jan. 2020 - Ghassen Chaabane, Moez Ezzeddine, Gaspard Donada--Vidal, Wafa Bouzouita

Summary

1. INTRODUCTION

What's Transfer learning? What's Knowledge Distillation?

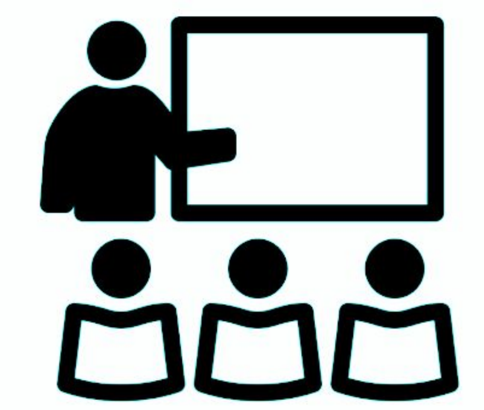
1. Introduction



1. Introduction

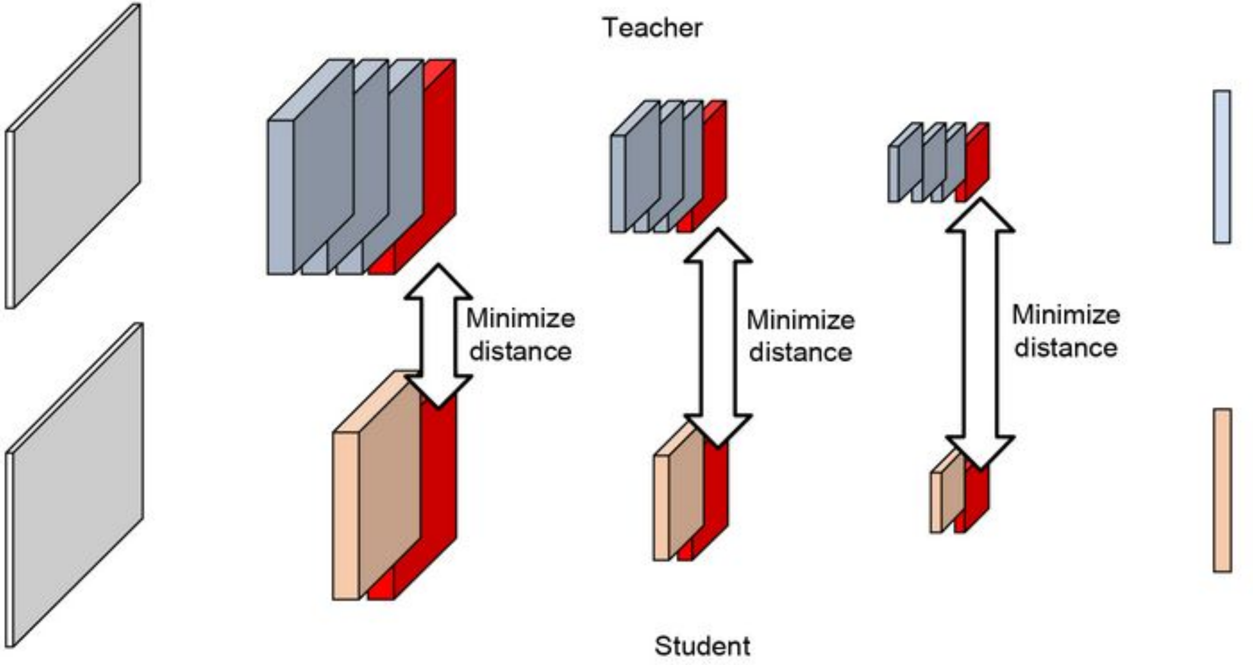
Knowledge Distillation

Distill a knowledge of
large and complex
network
(The teacher network)



Transfer it to a small and
simple network
(The student network)

1. Introduction



(This image was made by the authors of the article)

1. Introduction

1

```
Distillation epoch: 1
Train    Time Taken: 5.33 sec
layer1_activation similarity 66.4%
layer2_activation similarity 72.8%
layer3_activation similarity 62.1%

Distillation epoch: 2
Train    Time Taken: 4.95 sec
layer1_activation similarity 69.9%
layer2_activation similarity 76.1%
layer3_activation similarity 65.5%

Distillation epoch: 3
Train    Time Taken: 4.99 sec
layer1_activation similarity 74.7%
layer2_activation similarity 78.3%
```

2

```
Classification training Epoch: 1
Train    Time Taken: 4.28 sec
Loss: 0.319 | Acc: 90.280% (4514/5000)
Test     Time Taken: 2.91 sec
Loss: 0.293 | Acc: 91.670% (9167/10000)

Classification training Epoch: 2
Train    Time Taken: 4.32 sec
Loss: 0.060 | Acc: 99.080% (4954/5000)
Test     Time Taken: 2.72 sec
Loss: 0.244 | Acc: 92.950% (9295/10000)

Classification training Epoch: 3
Train    Time Taken: 4.31 sec
Loss: 0.048 | Acc: 99.700% (4985/5000)
```

2. PAPER OVERVIEW

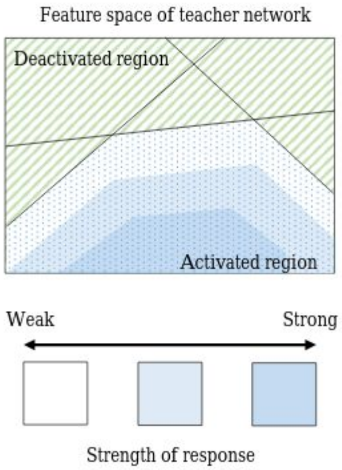
What’s the main idea?

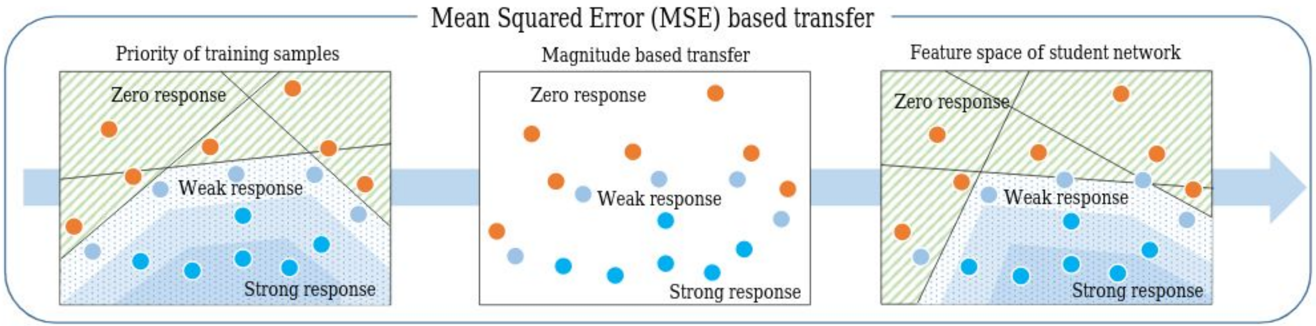
2. Paper Overview

- Seoul National University
  - Byeongho Heo
  - Minsik Lee
  - Sangdoo Yun
  - Jin Young Choi
- Presented in 2019
- Pretty well received by the community

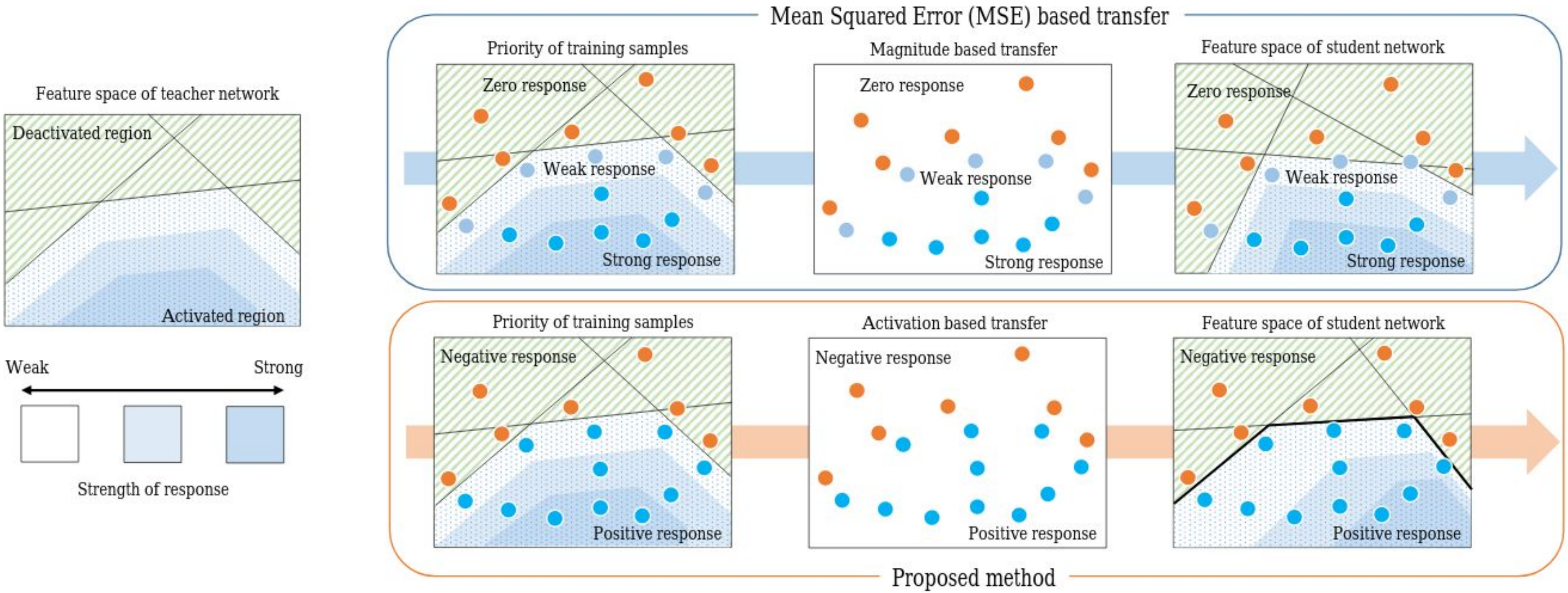
서울대학교
SEOUL NATIONAL UNIVERSITY

2. Paper Overview





2. Paper Overview



3. PROPOSED APPROACH

What’s the problem? What’s the proposal?

3.a Proposed approach - Activation boundaries

The activation boundary:

- separating hyperplane that determines whether neurons are active or deactivated.
- considered to be important for a long time.
- play an important role in forming the decision boundaries for classification-friendly partitioning of the feature space in each hidden layer.

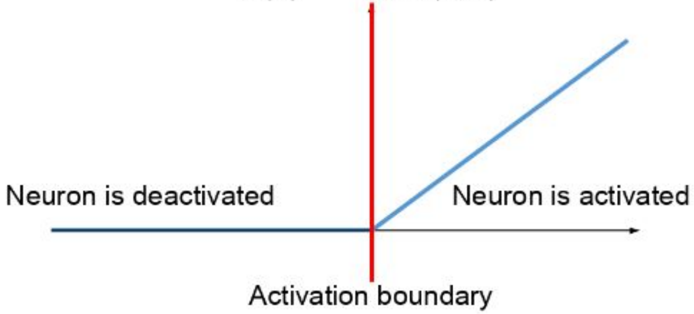
3.a Proposed approach - Activation boundaries

The activation boundary:

- separating hyperplane that determines whether neurons are active or deactivated.
- considered to be important for a long time.
- play an important role in forming the decision boundaries for classification-friendly partitioning of the feature space in each hidden layer.

ReLU example:

$$\sigma(x) = max(0, x)$$



3.b Proposed approach – The new method

Old studies and existing approach:

$$\mathcal{L}(\boldsymbol{I}) = \|\sigma(\mathcal{T}(\boldsymbol{I})) - \sigma(\mathcal{S}(\boldsymbol{I}))\|_2^2$$

(FITNET: Mean Squared Error based transfer)

3.b Proposed approach - The new method

Old studies and existing approach:

$$\mathcal{L}(\boldsymbol{I}) = \|\sigma(\mathcal{T}(\boldsymbol{I})) - \sigma(\mathcal{S}(\boldsymbol{I}))\|_2^2$$

Activation indicator function:

$$\rho(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{otherwise.} \end{cases}$$

3.b Proposed approach - The new method

Old studies and existing approach:

$$\mathcal{L}(\boldsymbol{I}) = \|\sigma(\mathcal{T}(\boldsymbol{I})) - \sigma(\mathcal{S}(\boldsymbol{I}))\|_2^2$$

Activation indicator function:

$$\rho(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Activation transfer loss:

$$\mathcal{L}(\boldsymbol{I}) = \|\rho(\mathcal{T}(\boldsymbol{I})) - \rho(\mathcal{S}(\boldsymbol{I}))\|_1$$

3.b Proposed approach - The new method

Old studies and existing approach:

$$\mathcal{L}(\boldsymbol{I}) = \|\sigma(\mathcal{T}(\boldsymbol{I})) - \sigma(\mathcal{S}(\boldsymbol{I}))\|_2^2$$

Activation indicator function:

$$\rho(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Activation transfer loss:

$$\mathcal{L}(\boldsymbol{I}) = \|\rho(\mathcal{T}(\boldsymbol{I})) - \rho(\mathcal{S}(\boldsymbol{I}))\|_1$$

Alternative loss:

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

3.b Proposed approach - The new method

Old studies and existing approach:

$$\mathcal{L}(\boldsymbol{I}) = \|\sigma(\mathcal{T}(\boldsymbol{I})) - \sigma(\mathcal{S}(\boldsymbol{I}))\|_2^2$$

Activation indicator function:

$$\rho(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Activation transfer loss:

$$\mathcal{L}(\boldsymbol{I}) = \|\rho(\mathcal{T}(\boldsymbol{I})) - \rho(\mathcal{S}(\boldsymbol{I}))\|_1$$

Alternative loss:

$$\mathcal{L}(\boldsymbol{I}) = \|\rho(\mathcal{T}(\boldsymbol{I})) \odot \underbrace{\boxed{\sigma(\mu\mathbf{1} - \mathcal{S}(\boldsymbol{I}))}}_{\text{Teacher neuron is activated}} + (\mathbf{1} - \rho(\mathcal{T}(\boldsymbol{I}))) \odot \underbrace{\boxed{\sigma(\mu\mathbf{1} + \mathcal{S}(\boldsymbol{I}))}}_{\text{Teacher neuron is deactivated}}\|_2^2$$

4. EXPERIMENTS

How did the author test his method? What are the results?

4. Experiments

Knowledge transfer for the same task

- When there is trained large network
- The goal is to train a small network that does the same task

Transfer learning

- Training a new network without pre-training
- Knowledge transfer can make similar effect of pre-training

Goals

- Fast training (Training epochs: distillation + classification)
- Little training data (Dataset’s size)
- Various network sizes (Type of compression)

4.a Experiments - Fast training

Less training epochs

Training epochs	1+1	1+5	3+12	5+25	10+50	20+100
Without distillation	43.37%	17.72%	11.76%	8.63%	7.41%	6.47%
Output distillation (KD)	48.42%	19.80%	12.09%	8.66%	6.80%	6.19%
FITNET + KD	48.16%	19.82%	11.10%	8.38%	7.02%	6.28%
FSP + KD	43.51%	19.29%	11.15%	8.48%	6.87%	6.22%
AT + KD	37.66%	14.14%	8.35%	6.68%	5.94%	5.80%
Jacobian + KD	38.46%	14.29%	8.37%	6.98%	5.98%	5.77%
Proposed + KD	**16.39%**	**11.16%**	**6.95%**	**6.08%**	**5.72%**	**5.58%**

Error rate

4.a Experiments – Fast training

Less training epochs ←

Training epochs	1+1	1+5	3+12	5+25	10+50	20+100
Without distillation	43.37%	17.72%	11.76%	8.63%	7.41%	6.47%
Output distillation (KD)	48.42%	19.80%	12.09%	8.66%	6.80%	6.19%
FITNET + KD	48.16%	19.82%	11.10%	8.38%	7.02%	6.28%
FSP + KD	43.51%	19.29%	11.15%	8.48%	6.87%	6.22%
AT + KD	37.66%	14.14%	8.35%	6.68%	5.94%	5.80%
Jacobian + KD	38.46%	14.29%	8.37%	6.98%	5.98%	5.77%
Proposed + KD	**16.39%**	**11.16%**	**6.95%**	**6.08%**	**5.72%**	**5.58%**
	19.17	10.89	7.22	6.30	5.92	5.77

Error rate

4.b Experiments - Little training data

Less training data

Percentage of data	0.1%	1%	5%	10%	20%	100%
Without distillation	73.83%	48.41%	28.12%	21.76%	15.26%	6.47%
Output distillation (KD)	74.62%	48.34%	28.13%	21.21%	14.62%	6.19%
FITNET + KD	74.81%	49.91%	27.28%	21.42%	14.52%	6.28%
FSP + KD	73.96%	47.98%	26.90%	20.80%	13.85%	6.22%
AT + KD	67.54%	37.11%	18.86%	15.61%	9.94%	5.80%
Jacobian + KD	68.65%	36.99%	18.34%	15.03%	9.83%	5.77%
Proposed + KD	**50.32%**	**21.54%**	**14.99%**	**13.09%**	**9.16%**	**5.58%**

4.b Experiments - Little training data

Less training data

Percentage of data	0.1%	1%	5%	10%	20%	100%
Without distillation	73.83%	48.41%	28.12%	21.76%	15.26%	6.47%
Output distillation (KD)	74.62%	48.34%	28.13%	21.21%	14.62%	6.19%
FITNET + KD	74.81%	49.91%	27.28%	21.42%	14.52%	6.28%
FSP + KD	73.96%	47.98%	26.90%	20.80%	13.85%	6.22%
AT + KD	67.54%	37.11%	18.86%	15.61%	9.94%	5.80%
Jacobian + KD	68.65%	36.99%	18.34%	15.03%	9.83%	5.77%
Proposed + KD	**50.32%**	**21.54%**	**14.99%**	**13.09%**	**9.16%**	**5.58%**
	-	20.61	13.44	12.15	8.84	5.77

4.c Experiments - Various network sizes

Compression type	Teacher	Student
Depth	WRN 22-4	WRN 10-4
Channel	WRN 16-4	WRN 16-2
Depth & Channel	WRN 22-4	WRN 16-2
Tiny network	WRN 22-4	WRN 10-1
Same network	WRN 16-4	WRN 16-4

The connector function: $r : \mathbb{R}^N \to \mathbb{R}^M$

M $(\mathcal{T}(\boldsymbol{I}) \in \mathbb{R}^M)$ The number of neurons in a teacher network

N $(\mathcal{S}(\boldsymbol{I}) \in \mathbb{R}^N)$ The number of neurons in a student network

- converts a neuron response vector of student to the size of teacher vector.

4.c Experiments - Various network sizes

Compression type	Teacher	Student
Depth	WRN 22-4	WRN 10-4
Channel	WRN 16-4	WRN 16-2
Depth & Channel	WRN 22-4	WRN 16-2
Tiny network	WRN 22-4	WRN 10-1
Same network	WRN 16-4	WRN 16-4

The connector function: $r : \mathbb{R}^N \to \mathbb{R}^M$

- Using the connector function, the alternative loss is changed as

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

4.c Experiments - Various network sizes

Compression type	Teacher	Student
Depth	WRN 22-4	WRN 10-4
Channel	WRN 16-4	WRN 16-2
Depth & Channel	WRN 22-4	WRN 16-2
Tiny network	WRN 22-4	WRN 10-1
Same network	WRN 16-4	WRN 16-4

The connector function: $r : \mathbb{R}^N \to \mathbb{R}^M$

- Using the connector function, the alternative loss is changed as

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

4.c Experiments - Various network sizes

Compression type	Teacher	Student
Depth	WRN 22-4	WRN 10-4
Channel	WRN 16-4	WRN 16-2
Depth & Channel	WRN 22-4	WRN 16-2
Tiny network	WRN 22-4	WRN 10-1
Same network	WRN 16-4	WRN 16-4

The connector function: $r : \mathbb{R}^N \to \mathbb{R}^M$

- Using the connector function, the alternative loss is changed as

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

Compression type	Size ratio	KD	FITNET	FSP	AT	Jacobian	Proposed
Depth	27.9%	22.98%	23.34%	22.99%	18.06%	18.28%	**14.05%**
Channel	25.2%	20.48%	19.98%	19.78%	14.81%	14.41%	**11.62%**
Depth & Channel	16.1%	21.21%	21.42%	20.80%	15.61%	15.03%	**13.09%**
Tiny network	1.8%	29.57%	29.18%	28.70%	29.44%	28.70%	**23.27%**
Same network	100%	18.29%	17.91%	17.81%	12.03%	11.28%	**6.63%**

4.c Experiments - Various network sizes

Compression type	Teacher	Student
Depth	WRN 22-4	WRN 10-4
Channel	WRN 16-4	WRN 16-2
Depth & Channel	WRN 22-4	WRN 16-2
Tiny network	WRN 22-4	WRN 10-1
Same network	WRN 16-4	WRN 16-4

The connector function: $r : \mathbb{R}^N \rightarrow \mathbb{R}^M$

- Using the connector function, the alternative loss is changed as

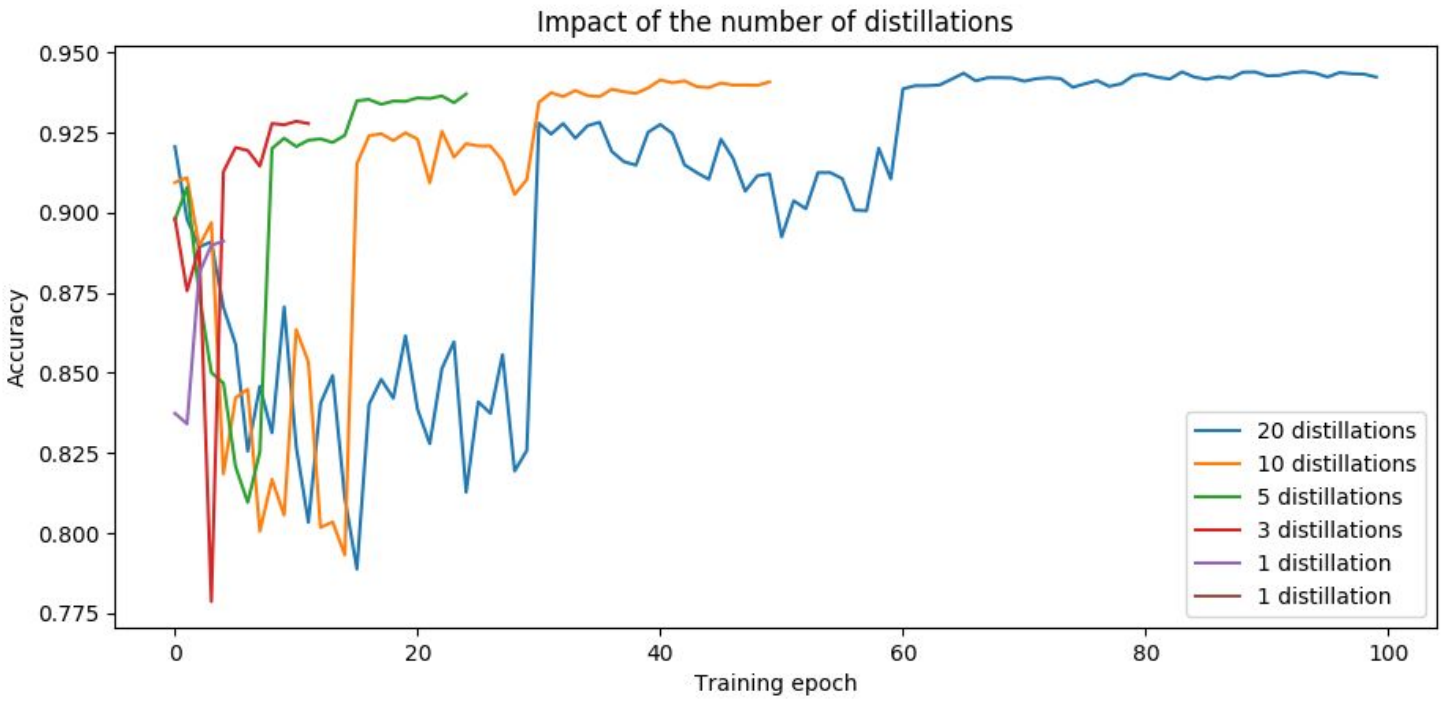
$$\mathcal{L}(\boldsymbol{I}) = \|\rho(\mathcal{T}(\boldsymbol{I})) \odot \sigma(\mu \mathbf{1} - r(\mathcal{S}(\boldsymbol{I}))) + (\mathbf{1} - \rho(\mathcal{T}(\boldsymbol{I}))) \odot \sigma(\mu \mathbf{1} + r(\mathcal{S}(\boldsymbol{I})))\|_2^2.$$

Compression type	Size ratio	KD	FITNET	FSP	AT	Jacobian	Proposed	
Depth	27.9%	22.98%	23.34%	22.99%	18.06%	18.28%	**14.05%**	13.58
Channel	25.2%	20.48%	19.98%	19.78%	14.81%	14.41%	**11.62%**	11.43
Depth & Channel	16.1%	21.21%	21.42%	20.80%	15.61%	15.03%	**13.09%**	12.15
Tiny network	1.8%	29.57%	29.18%	28.70%	29.44%	28.70%	**23.27%**	22.30
Same network	100%	18.29%	17.91%	17.81%	12.03%	11.28%	**6.63%**	6.01

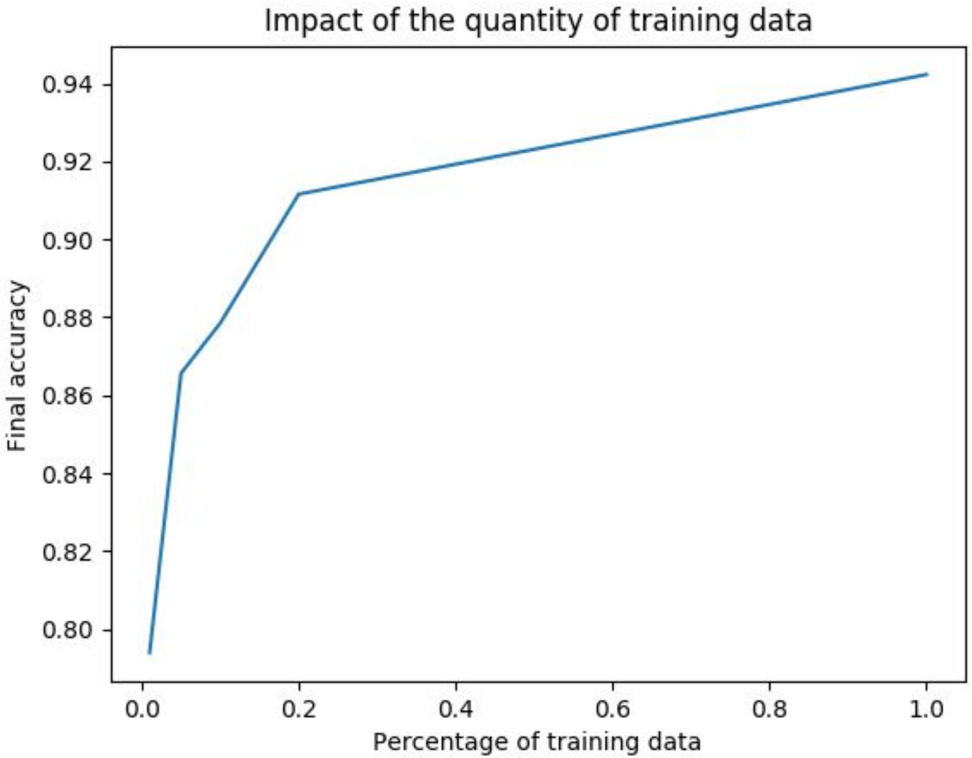
5. DISCUSSION

What are the study limitations?

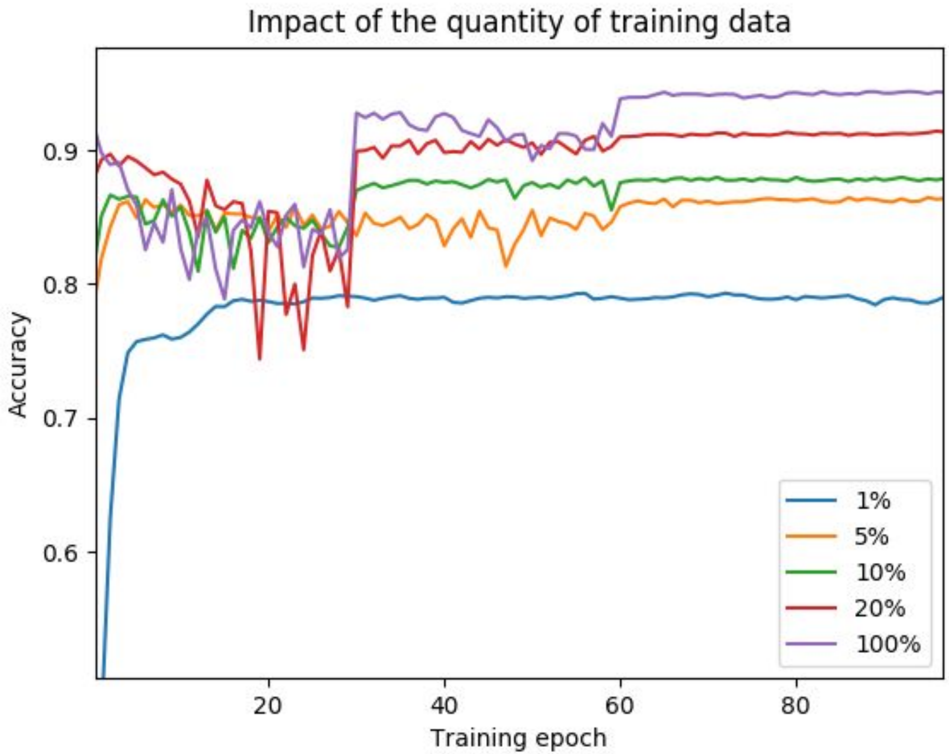
5. Discussion



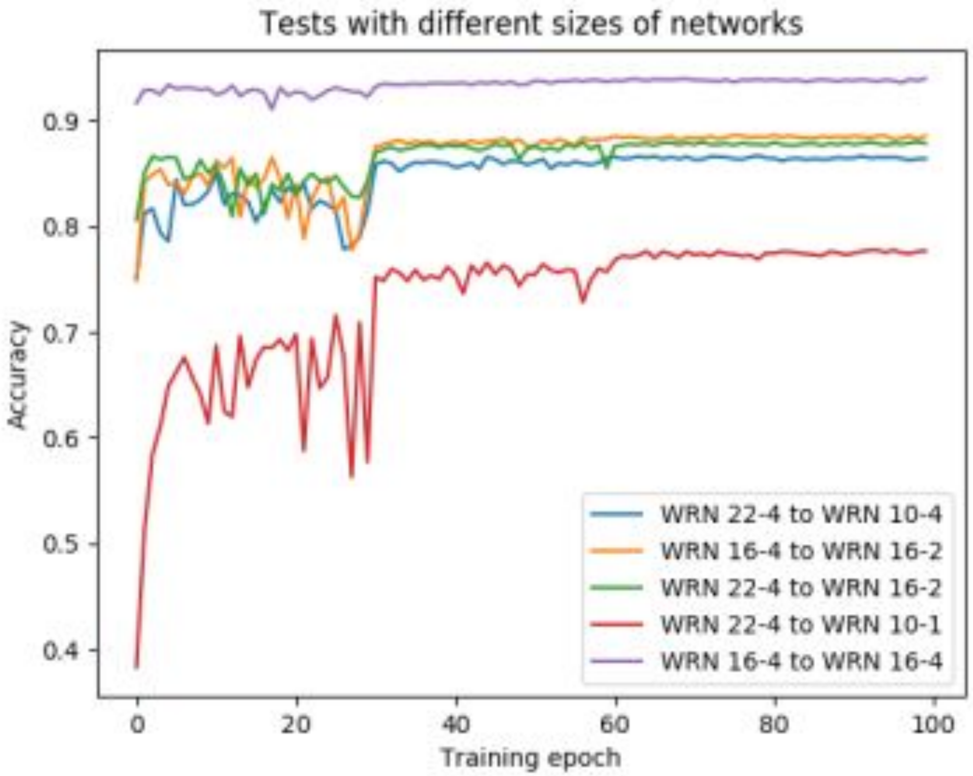
5. Discussion



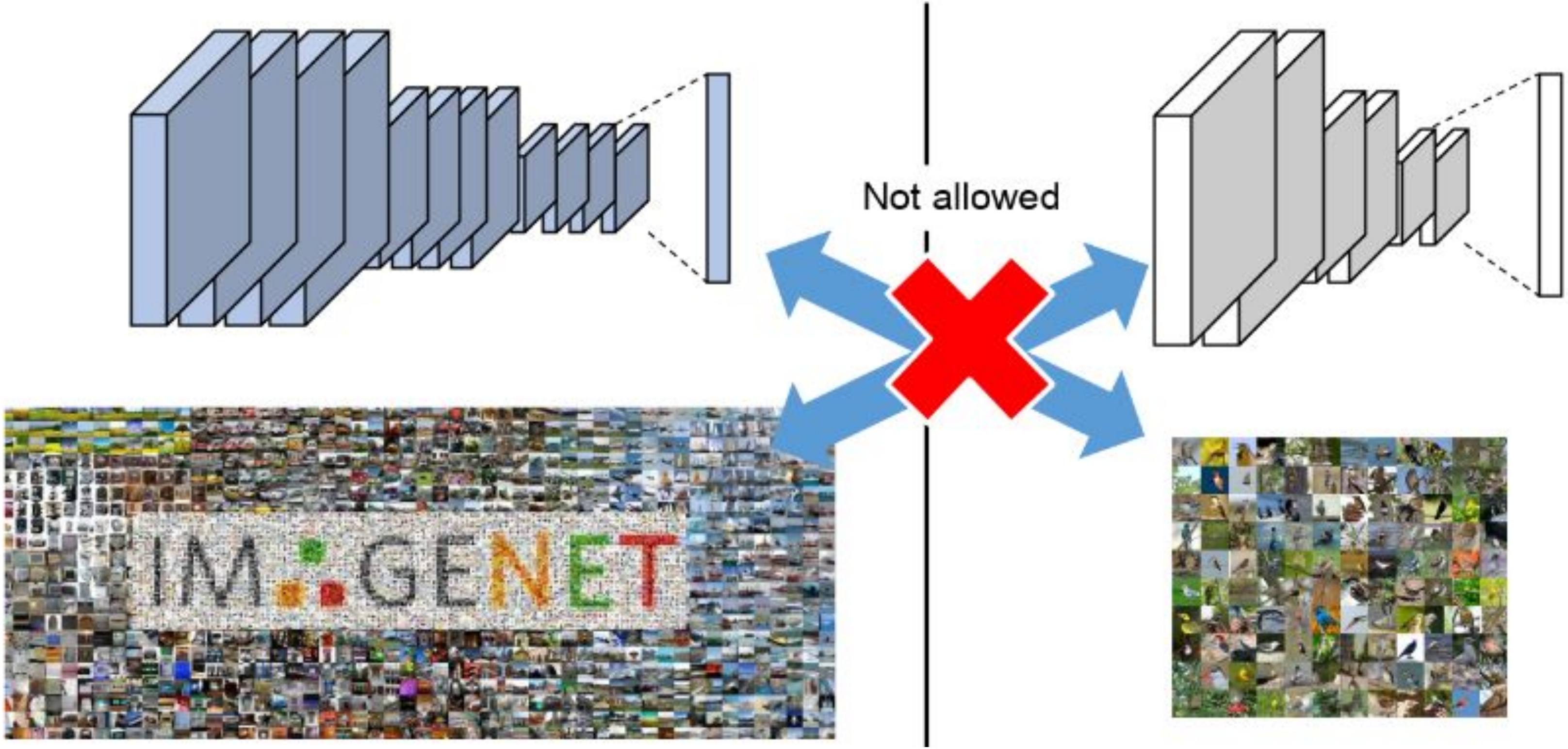
5. Discussion



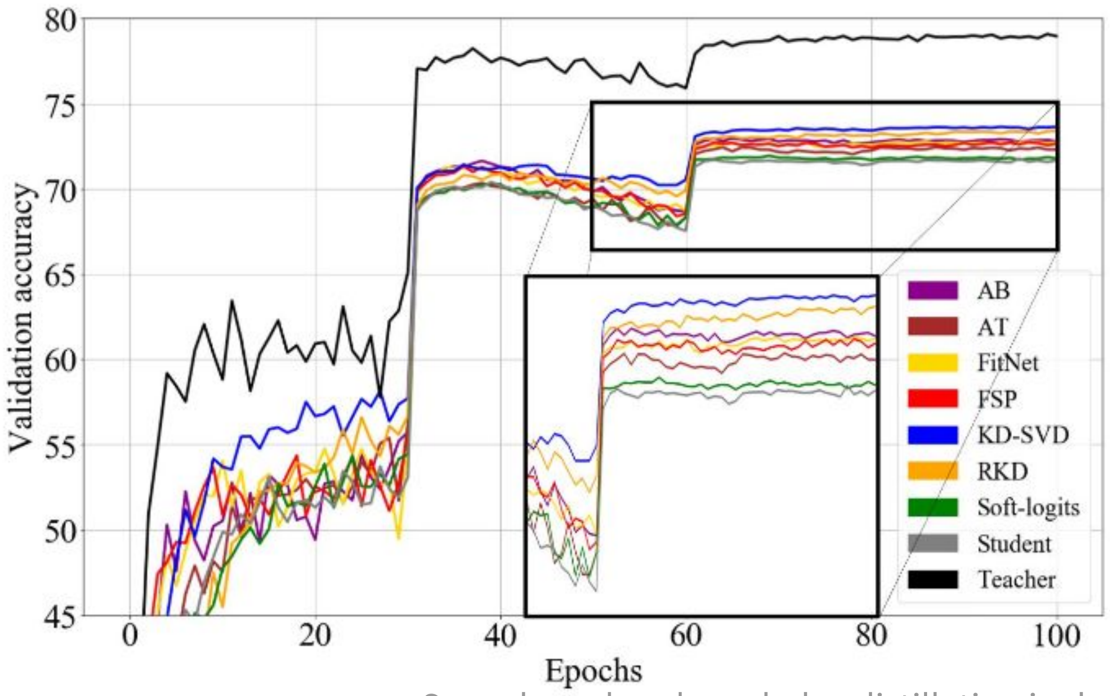
5. Discussion



5. Discussion



5. Discussion



Seunghyun lee : knowledge distillation in deep neural network

5. Discussion

Pros	Cons
The main idea and the proposed approach	
The accuracy and precision in the mentioned results	The variety of experiments in transfer learning section
Explanations’ clarity and rigor	

5. Discussion

Pros	Cons
The main idea and the proposed approach	
The accuracy and precision in the mentioned results	The variety of experiments in transfer learning section
Explanations’ clarity and rigor	

6. CONCLUSION

What did we learn?

6. Conclusion

- 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

6. Conclusion

- 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: defining the knowledge

Thank you for your attention!

”