Sequential Knowledge Transfer in Teacher-Student Framework Using Densely Distilled Flow-Based Information

Electronics and Telecommunications Research Institute

ICIP 2018

Doyeob Yeo, Ji-Hoon Bae, Junho Yim, Nae-Soo Kim, Cheol-Sig Pyo, and Junmo Kim

ETRI
Electronics and Telecommunications Research Institute

KAIST

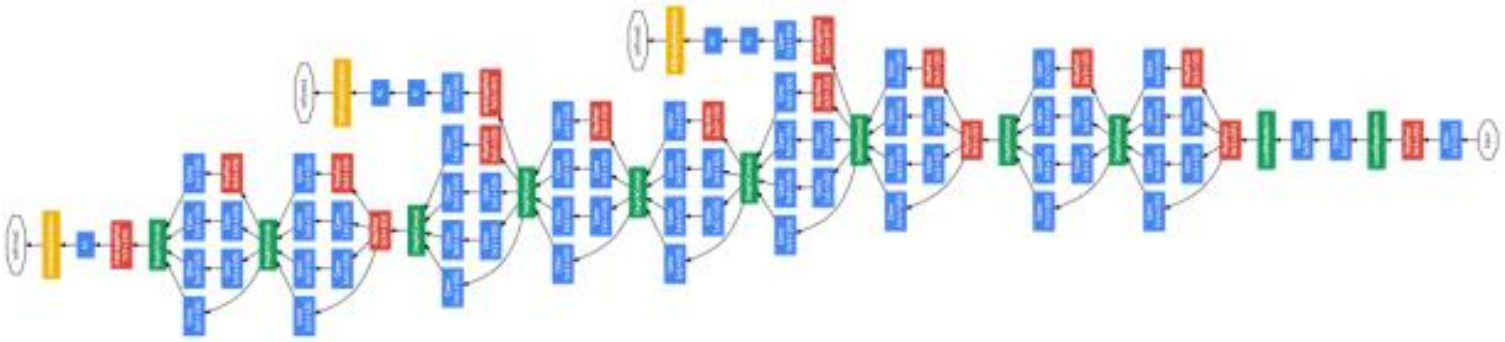
Contents

- Introduction & Motivation
- Previous Research
- Proposed Model
- Experimental Results
- Conclusion

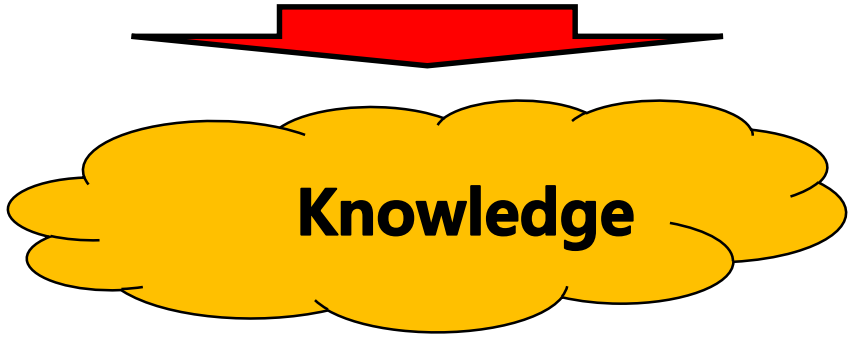
Introduction

- Knowledge Distillation

Deep Neural Network



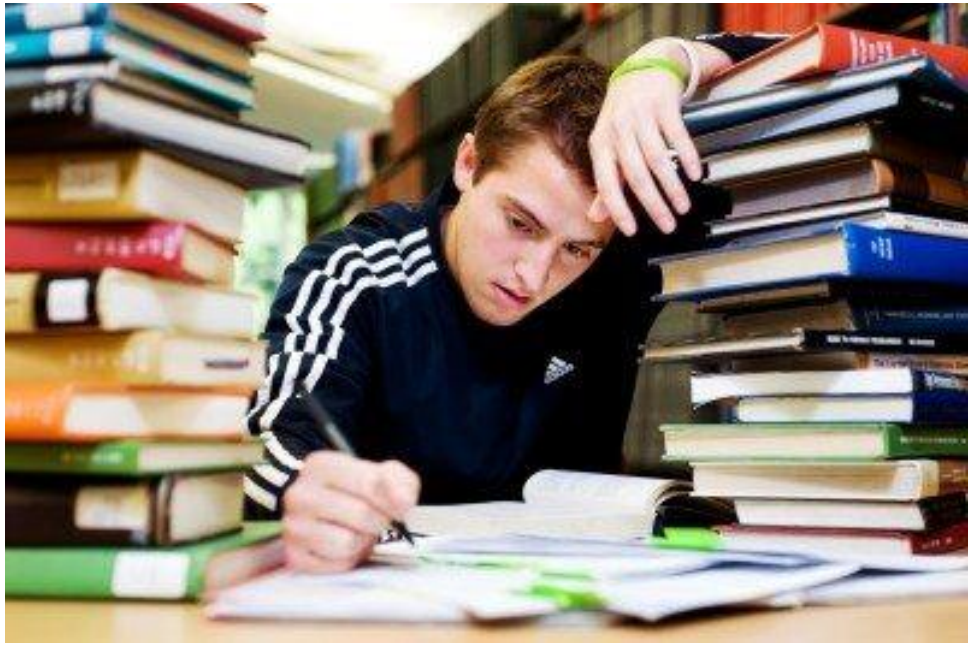
Extract the useful Knowledge



The Google "Inception" deep neural network architecture. Source: Christian Szegedy et. al. Going deeper with convolutions. CVPR 2015.

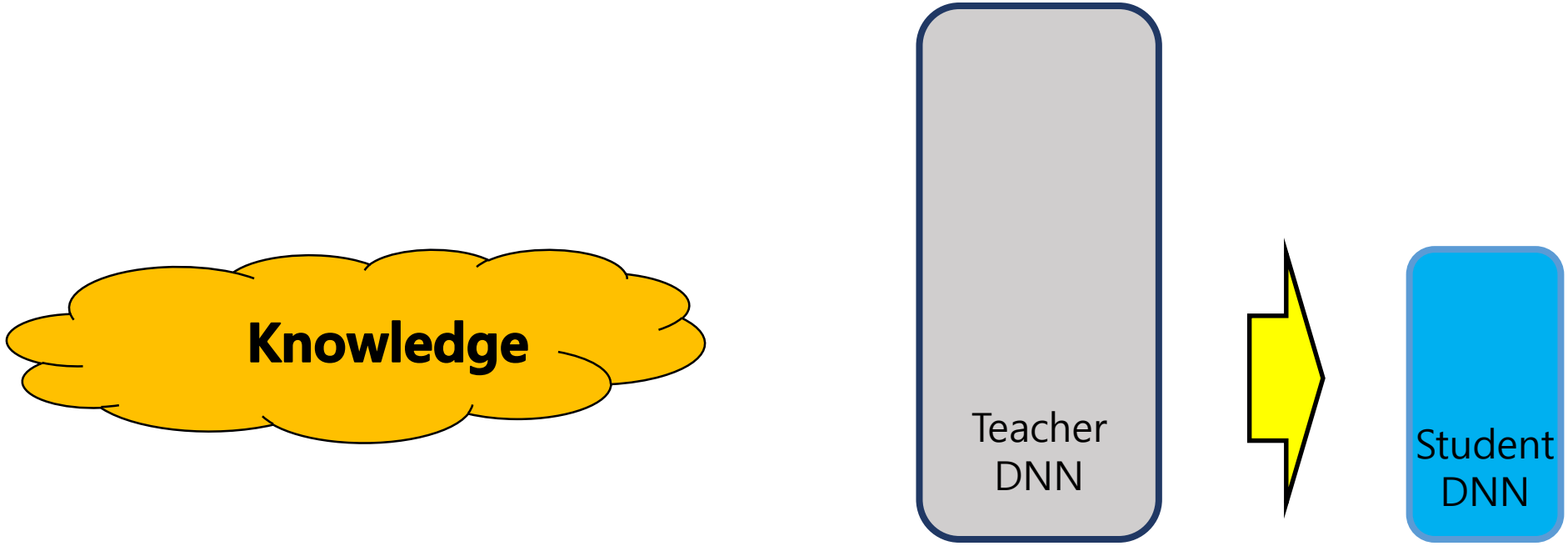
Motivation

- Knowledge Distillation with Teacher – Student framework



Motivation

- Two main issues in knowledge distillation method



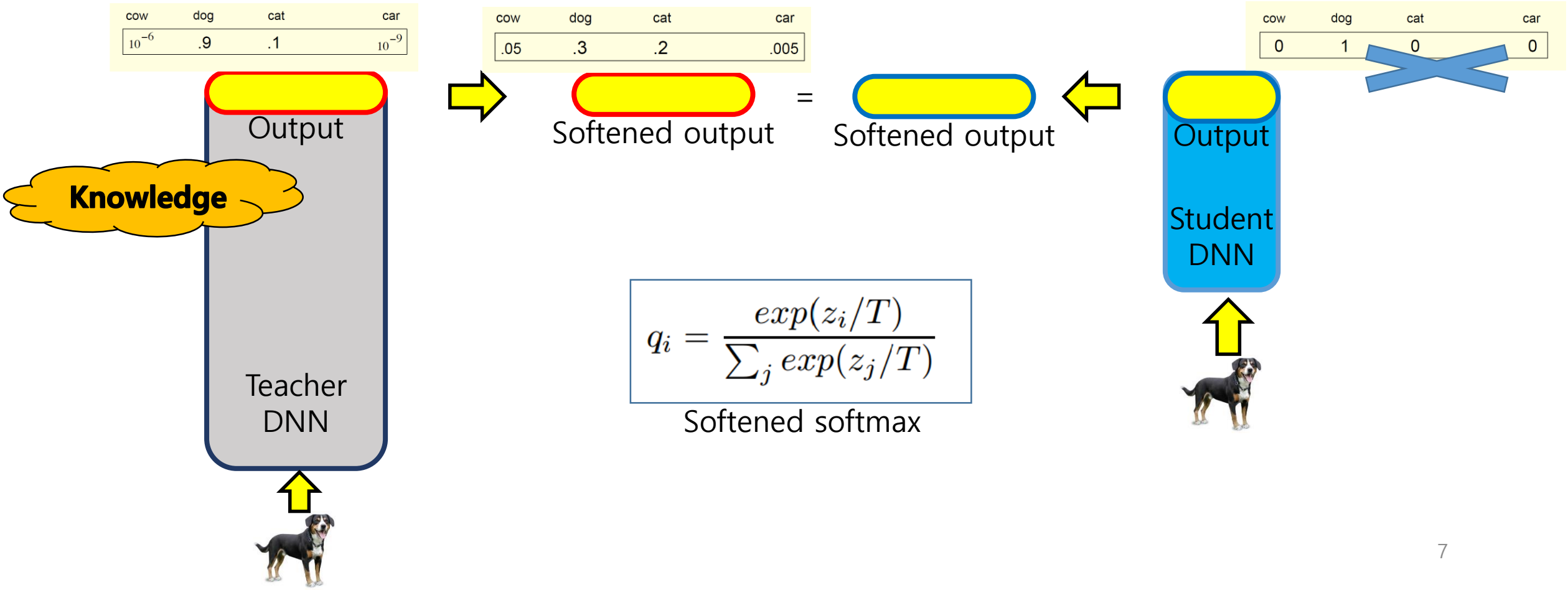
What kind of Knowledge?

How to transfer the knowledge?

Previous Research

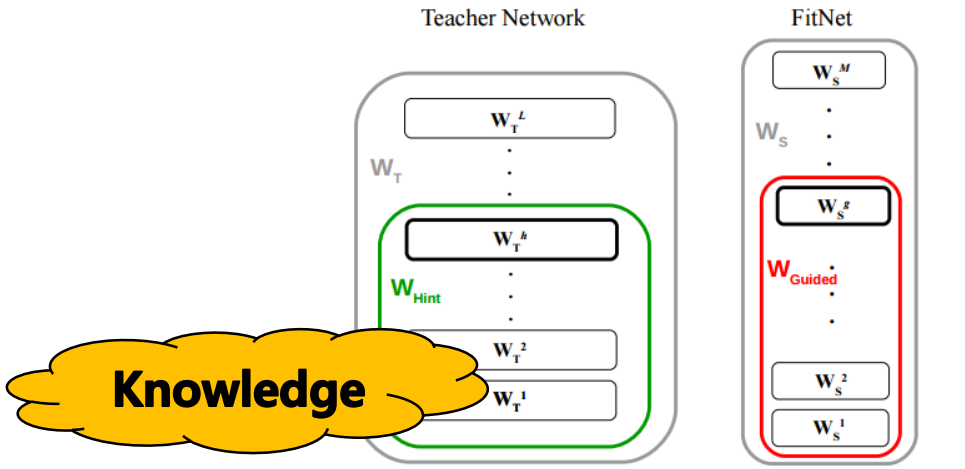
Previous Research Knowledge Distillation

- Hinton et al. "Distilling the knowledge in a neural network", arXiv 2014
  - Student DNN is penalized according to a softened version of the teacher DNN's output

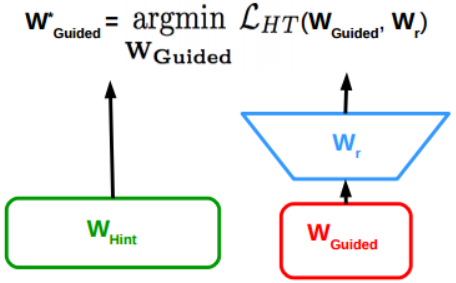


Previous Research Knowledge Distillation

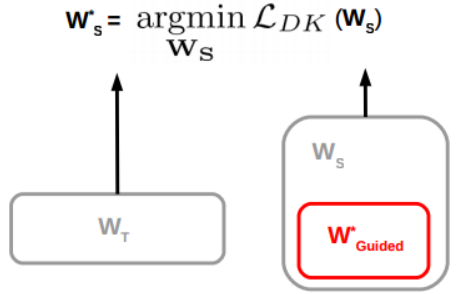
- Romero et al. “Fitnets: Hints for thin deep nets”, ICLR 2015
  - Student DNN is also penalized according to a intermediate features of the teacher DNN



(a) Teacher and Student Networks



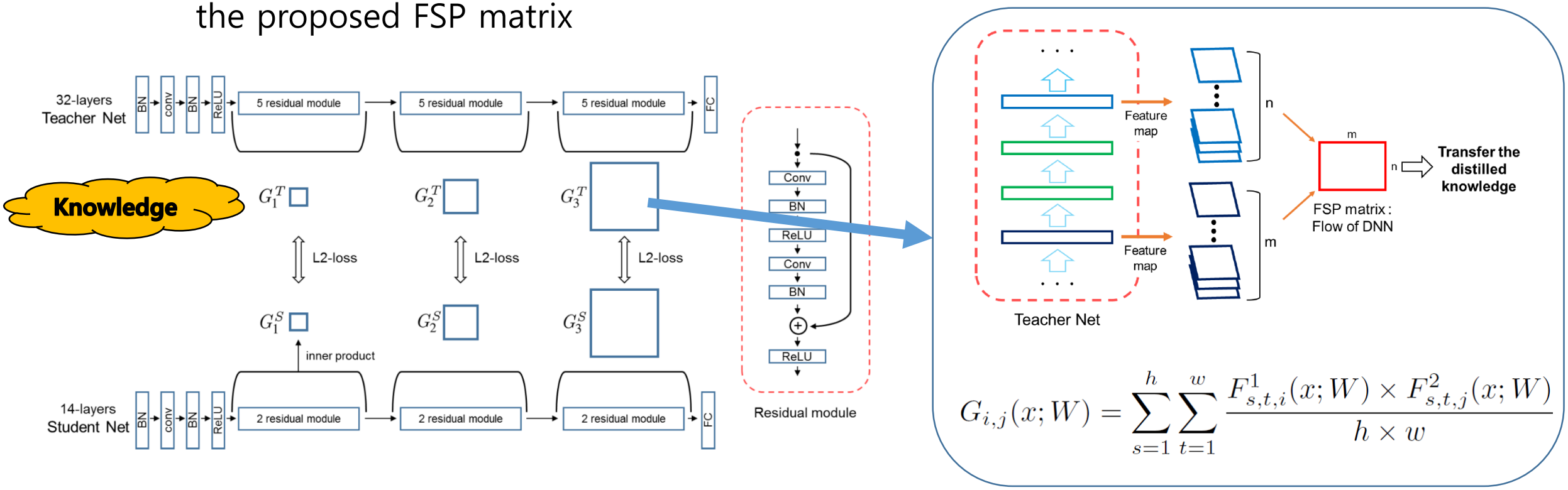
(b) Hints Training



(c) Knowledge Distillation

Previous Research Knowledge Distillation

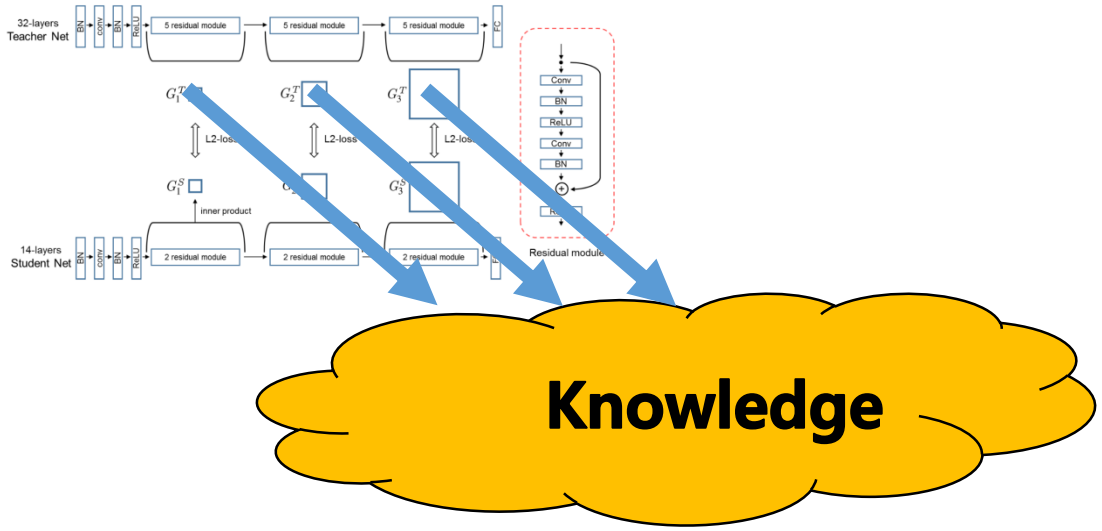
- Yim et al. “Gift from Knowledge distillation: Fast Optimization, Network Minimization and Transfer Learning”, CVPR 2017
  - Determine the distilled knowledge as the flow of the solving procedure calculated with the proposed FSP matrix

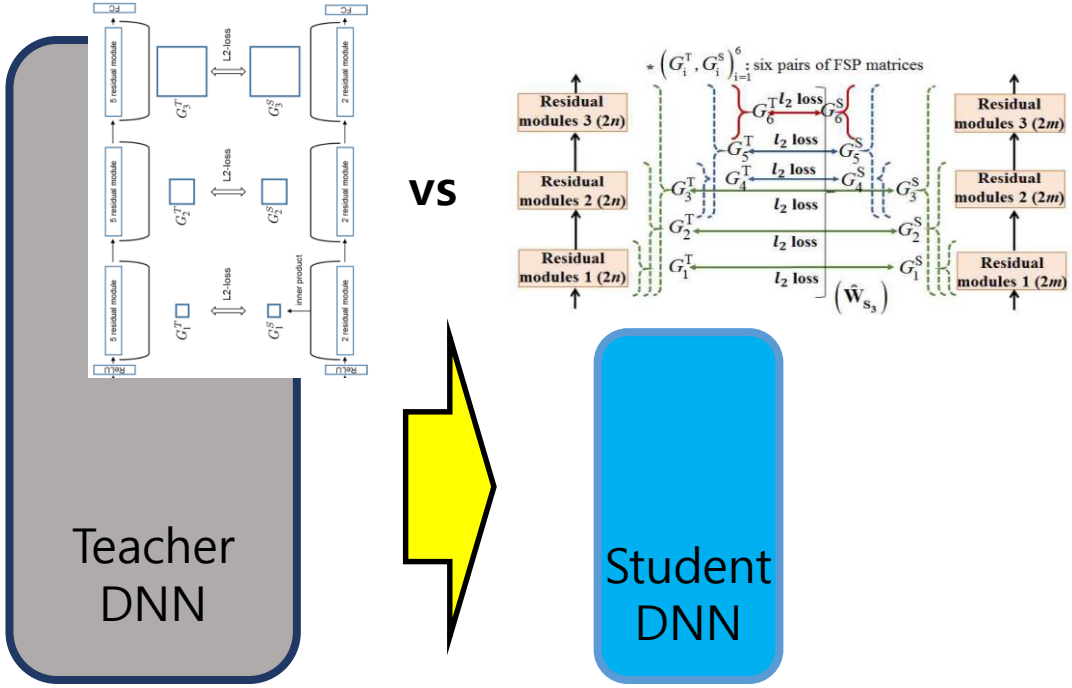


Proposed Model

Motivation

- Two main issues in knowledge distillation method





What kind of Knowledge?

How to transfer the knowledge?

Proposed Model Sequential Knowledge Transfer

- Part of Table of contents of Thomas' Calculus

2. Limits and Continuity
2.1 Rates of Change and Tangents to Curves
2.2 Limit of a Function and Limit Laws
2.3 The Precise Definition of a Limit
2.4 One-Sided Limits
2.5 Continuity
2.6 Limits Involving Infinity; Asymptotes of Graphs

3. Differentiation
3.1 Tangents and the Derivative at a Point
3.2 The Derivative as a Function
3.3 Differentiation Rules
3.4 The Derivative as a Rate of Change
3.5 Derivatives of Trigonometric Functions
3.6 The Chain Rule
3.7 Implicit Differentiation
3.8 Related Rates
3.9 Linearization and Differentials

4. Applications of Derivatives
4.1 Extreme Values of Functions
4.2 The Mean Value Theorem
4.3 Monotonic Functions and the First Derivative Test
4.4 Concavity and Curve Sketching
4.5 Applied Optimization
4.6 Newton's Method
4.7 Antiderivatives

http://www.mypearsonstore.com/bookstore/thomas-calculus-9780321587992

Proposed Model Sequential Knowledge Transfer

- Part of Table of contents of Thomas' Calculus

2. Limits and Continuity
2.1 Rates of Change and Tangents to Curves
2.2 Limit of a Function and Limit Laws
2.3 The Precise Definition of a Limit
2.4 One-Sided Limits
2.5 Continuity
2.6 Limits Involving Infinity; Asymptotes of Graphs

3. Differentiation
3.1 Tangents and the Derivative at a Point
3.2 The Derivative as a Function
3.3 Differentiation Rules
3.4 The Derivative as a Rate of Change
3.5 Derivatives of Trigonometric Functions
3.6 The Chain Rule
3.7 Implicit Differentiation
3.8 Related Rates
3.9 Linearization and Differentials

4. Applications of Derivatives
4.1 Extreme Values of Functions
4.2 The Mean Value Theorem
4.3 Monotonic Functions and the First Derivative Test
4.4 Concavity and Curve Sketching
4.5 Applied Optimization
4.6 Newton's Method
4.7 Antiderivatives

Proposed Model Sequential Knowledge Transfer

- Part of Table of contents of Thomas' Calculus

2. Limits and Continuity
2.1 Rates of Change and Tangents to Curves
2.2 Limit of a Function and Limit Laws
2.3 The Precise Definition of a Limit
2.4 One-Sided Limits
2.5 Continuity
2.6 Limits Involving Infinity; Asymptotes of Graphs

3. Differentiation
3.1 Tangents and the Derivative at a Point
3.2 The Derivative as a Function
3.3 Differentiation Rules
3.4 The Derivative as a Rate of Change
3.5 Derivatives of Trigonometric Functions
3.6 The Chain Rule
3.7 Implicit Differentiation
3.8 Related Rates
3.9 Linearization and Differentials

4. Applications of Derivatives
4.1 Extreme Values of Functions
4.2 The Mean Value Theorem
4.3 Monotonic Functions and the First Derivative Test
4.4 Concavity and Curve Sketching
4.5 Applied Optimization
4.6 Newton's Method
4.7 Antiderivatives

Proposed Model Sequential Knowledge Transfer

- Part of Table of contents of Thomas' Calculus

2. Limits and Continuity
2.1 Rates of Change and Tangents to Curves
2.2 Limit of a Function and Limit Laws
2.3 The Precise Definition of a Limit
2.4 One-Sided Limits
2.5 Continuity
2.6 Limits Involving Infinity; Asymptotes of Graphs

3. Differentiation
3.1 Tangents and the Derivative at a Point
3.2 The Derivative as a Function
3.3 Differentiation Rules
3.4 The Derivative as a Rate of Change
3.5 Derivatives of Trigonometric Functions
3.6 The Chain Rule
3.7 Implicit Differentiation
3.8 Related Rates
3.9 Linearization and Differentials

4. Applications of Derivatives
4.1 Extreme Values of Functions
4.2 The Mean Value Theorem
4.3 Monotonic Functions and the First Derivative Test
4.4 Concavity and Curve Sketching
4.5 Applied Optimization
4.6 Newton's Method
4.7 Antiderivatives

Proposed Model Sequential Knowledge Transfer

- Part of Table of contents of Thomas' Calculus

2. Limits and Continuity
2.1 Rates of Change and Tangents to Curves
2.2 Limit of a Function and Limit Laws
2.3 The Precise Definition of a Limit
2.4 One-Sided Limits
2.5 Continuity
2.6 Limits Involving Infinity; Asymptotes of Graphs

3. Differentiation
3.1 Tangents and the Derivative at a Point
3.2 The Derivative as a Function
3.3 Differentiation Rules
3.4 The Derivative as a Rate of Change
3.5 Derivatives of Trigonometric Functions
3.6 The Chain Rule
3.7 Implicit Differentiation
3.8 Related Rates
3.9 Linearization and Differentials

4. Applications of Derivatives
4.1 Extreme Values of Functions
4.2 The Mean Value Theorem
4.3 Monotonic Functions and the First Derivative Test
4.4 Concavity and Curve Sketching
4.5 Applied Optimization
4.6 Newton's Method
4.7 Antiderivatives

Proposed Model Sequential Knowledge Transfer

- Part of Table of contents of Thomas' Calculus

2. Limits and Continuity
2.1 Rates of Change and Tangents to Curves
2.2 Limit of a Function and Limit Laws
2.3 The Precise Definition of a Limit
2.4 One-Sided Limits
2.5 Continuity
2.6 Limits Involving Infinity; Asymptotes of Graphs

3. Differentiation
3.1 Tangents and the Derivative at a Point
3.2 The Derivative as a Function
3.3 Differentiation Rules
3.4 The Derivative as a Rate of Change
3.5 Derivatives of Trigonometric Functions
3.6 The Chain Rule
3.7 Implicit Differentiation
3.8 Related Rates
3.9 Linearization and Differentials

4. Applications of Derivatives
4.1 Extreme Values of Functions
4.2 The Mean Value Theorem
4.3 Monotonic Functions and the First Derivative Test
4.4 Concavity and Curve Sketching
4.5 Applied Optimization
4.6 Newton's Method
4.7 Antiderivatives

Proposed Model Sequential Knowledge Transfer

- Part of Table of contents of Thomas' Calculus

2. Limits and Continuity
2.1 Rates of Change and Tangents to Curves
2.2 Limit of a Function and Limit Laws
2.3 The Precise Definition of a Limit
2.4 One-Sided Limits
2.5 Continuity
2.6 Limits Involving Infinity; Asymptotes of Graphs

3. Differentiation
3.1 Tangents and the Derivative at a Point
3.2 The Derivative as a Function
3.3 Differentiation Rules
3.4 The Derivative as a Rate of Change
3.5 Derivatives of Trigonometric Functions
3.6 The Chain Rule
3.7 Implicit Differentiation
3.8 Related Rates
3.9 Linearization and Differentials

4. Applications of Derivatives
4.1 Extreme Values of Functions
4.2 The Mean Value Theorem
4.3 Monotonic Functions and the First Derivative Test
4.4 Concavity and Curve Sketching
4.5 Applied Optimization
4.6 Newton's Method
4.7 Antiderivatives

Proposed Model Sequential Knowledge Transfer

- Step 1

2. Limits and Continuity
2.1 Rates of Change and Tangents to Curves
2.2 Limit of a Function and Limit Laws
2.3 The Precise Definition of a Limit
2.4 One-Sided Limits
2.5 Continuity
2.6 Limits Involving Infinity; Asymptotes of Graphs

3. Differentiation
3.1 Tangents and the Derivative at a Point
3.2 The Derivative as a Function
3.3 Differentiation Rules
3.4 The Derivative as a Rate of Change
3.5 Derivatives of Trigonometric Functions
3.6 The Chain Rule
3.7 Implicit Differentiation
3.8 Related Rates
3.9 Linearization and Differentials

4. Applications of Derivatives
4.1 Extreme Values of Functions
4.2 The Mean Value Theorem
4.3 Monotonic Functions and the First Derivative Test
4.4 Concavity and Curve Sketching
4.5 Applied Optimization
4.6 Newton's Method
4.7 Antiderivatives

2. Limits and Continuity
2.1 Rates of Change and Tangents to Curves
2.2 Limit of a Function and Limit Laws
2.3 The Precise Definition of a Limit
2.4 One-Sided Limits
2.5 Continuity
2.6 Limits Involving Infinity; Asymptotes of Graphs

3. Differentiation
3.1 Tangents and the Derivative at a Point
3.2 The Derivative as a Function
3.3 Differentiation Rules
3.4 The Derivative as a Rate of Change
3.5 Derivatives of Trigonometric Functions
3.6 The Chain Rule
3.7 Implicit Differentiation
3.8 Related Rates
3.9 Linearization and Differentials

4. Applications of Derivatives
4.1 Extreme Values of Functions
4.2 The Mean Value Theorem
4.3 Monotonic Functions and the First Derivative Test
4.4 Concavity and Curve Sketching
4.5 Applied Optimization
4.6 Newton's Method
4.7 Antiderivatives

2. Limits and Continuity
2.1 Rates of Change and Tangents to Curves
2.2 Limit of a Function and Limit Laws
2.3 The Precise Definition of a Limit
2.4 One-Sided Limits
2.5 Continuity
2.6 Limits Involving Infinity; Asymptotes of Graphs

3. Differentiation
3.1 Tangents and the Derivative at a Point
3.2 The Derivative as a Function
3.3 Differentiation Rules
3.4 The Derivative as a Rate of Change
3.5 Derivatives of Trigonometric Functions
3.6 The Chain Rule
3.7 Implicit Differentiation
3.8 Related Rates
3.9 Linearization and Differentials

4. Applications of Derivatives
4.1 Extreme Values of Functions
4.2 The Mean Value Theorem
4.3 Monotonic Functions and the First Derivative Test
4.4 Concavity and Curve Sketching
4.5 Applied Optimization
4.6 Newton's Method
4.7 Antiderivatives

Proposed Model Sequential Knowledge Transfer

- Step 1

2. Limits and Continuity
2.1 Rates of Change and Tangents to Curves
2.2 Limit of a Function and Limit Laws
2.3 The Precise Definition of a Limit
2.4 One-Sided Limits
2.5 Continuity
2.6 Limits Involving Infinity; Asymptotes of Graphs

3. Differentiation
3.1 Tangents and the Derivative at a Point
3.2 The Derivative as a Function
3.3 Differentiation Rules
3.4 The Derivative as a Rate of Change
3.5 Derivatives of Trigonometric Functions
3.6 The Chain Rule
3.7 Implicit Differentiation
3.8 Related Rates
3.9 Linearization and Differentials

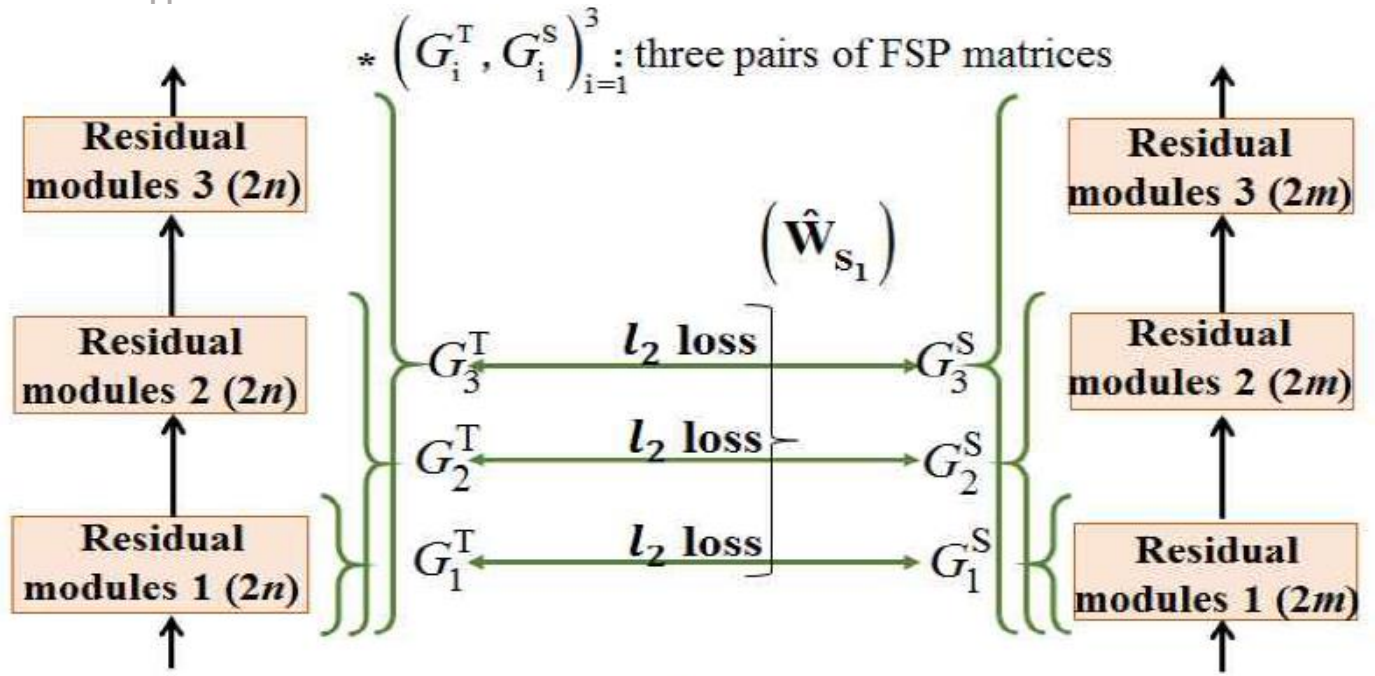
4. Applications of Derivatives

2. Limits and Continuity
2.1 Rates of Change and Tangents to Curves
2.2 Limit of a Function and Limit Laws
2.3 The Precise Definition of a Limit
2.4 One-Sided Limits
2.5 Continuity
2.6 Limits Involving Infinity; Asymptotes of Graphs

3. Differentiation
3.1 Tangents and the Derivative at a Point
3.2 The Derivative as a Function
3.3 Differentiation Rules
3.4 The Derivative as a Rate of Change
3.5 Derivatives of Trigonometric Functions
3.6 The Chain Rule
3.7 Implicit Differentiation
3.8 Related Rates
3.9 Linearization and Differentials

2. Limits and Continuity
2.1 Rates of Change and Tangents to Curves
2.2 Limit of a Function and Limit Laws
2.3 The Precise Definition of a Limit
2.4 One-Sided Limits
2.5 Continuity
2.6 Limits Involving Infinity; Asymptotes of Graphs

3. Differentiation
3.1 Tangents and the Derivative at a Point
3.2 The Derivative as a Function
3.3 Differentiation Rules
3.4 The Derivative as a Rate of Change
3.5 Derivatives of Trigonometric Functions
3.6 The Chain Rule
3.7 Implicit Differentiation
3.8 Related Rates
3.9 Linearization and Differentials



(b)

$$(\hat{\mathbf{W}}_{S_1}) = \arg\min_{\mathbf{W}_{S_1}} \sum_{i=1}^{n_0} \lambda_i \left\| G_i^T(x, \mathbf{W}_T) - G_i^S(x, \mathbf{W}_{S_1}) \right\|_2^2, \quad (1)$$

Proposed Model Sequential Knowledge Transfer

- Step 2

2. Limits and Continuity
2.1 Rates of Change and Tangents to Curves
2.2 Limit of a Function and Limit Laws
2.3 The Precise Definition of a Limit
2.4 One-Sided Limits
2.5 Continuity
2.6 Limits Involving Infinity; Asymptotes of Graphs

3. Differentiation
3.1 Tangents and the Derivative at a Point
3.2 The Derivative as a Function
3.3 Differentiation Rules
3.4 The Derivative as a Rate of Change
3.5 Derivatives of Trigonometric Functions
3.6 The Chain Rule
3.7 Implicit Differentiation
3.8 Related Rates
3.9 Linearization and Differentials

4. Applications of Derivatives
4.1 Extreme Values of Functions
4.2 The Mean Value Theorem
4.3 Monotonic Functions and the First Derivative Test
4.4 Concavity and Curve Sketching
4.5 Applied Optimization
4.6 Newton's Method
4.7 Antiderivatives

2. Limits and Continuity
2.1 Rates of Change and Tangents to Curves
2.2 Limit of a Function and Limit Laws
2.3 The Precise Definition of a Limit
2.4 One-Sided Limits
2.5 Continuity
2.6 Limits Involving Infinity; Asymptotes of Graphs

3. Differentiation
3.1 Tangents and the Derivative at a Point
3.2 The Derivative as a Function
3.3 Differentiation Rules
3.4 The Derivative as a Rate of Change
3.5 Derivatives of Trigonometric Functions
3.6 The Chain Rule
3.7 Implicit Differentiation
3.8 Related Rates
3.9 Linearization and Differentials

4. Applications of Derivatives
4.1 Extreme Values of Functions
4.2 The Mean Value Theorem
4.3 Monotonic Functions and the First Derivative Test
4.4 Concavity and Curve Sketching
4.5 Applied Optimization
4.6 Newton's Method
4.7 Antiderivatives

Proposed Model Sequential Knowledge Transfer

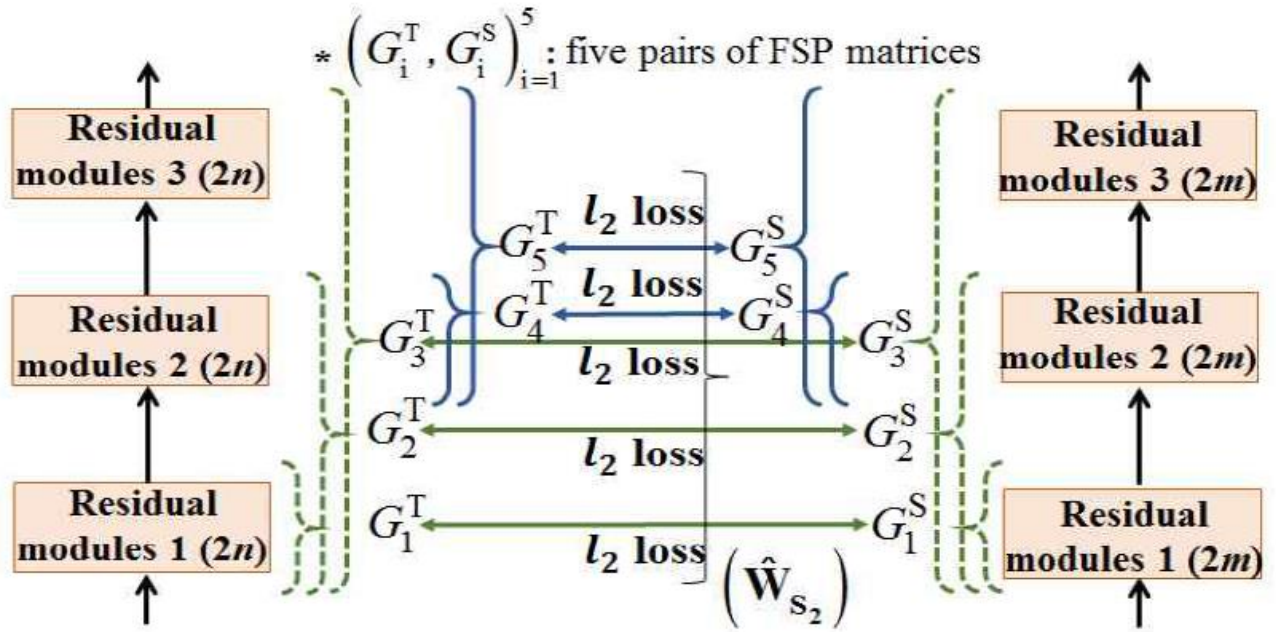
- Step 2

2. Limits and Continuity
2.1 Rates of Change and Tangents to Curves
2.2 Limit of a Function and Limit Laws
2.3 The Precise Definition of a Limit
2.4 One-Sided Limits
2.5 Continuity
2.6 Limits Involving Infinity; Asymptotes of Graphs

3. Differentiation
3.1 Tangents and the Derivative at a Point
3.2 The Derivative as a Function
3.3 Differentiation Rules
3.4 The Derivative as a Rate of Change
3.5 Derivatives of Trigonometric Functions
3.6 The Chain Rule
3.7 Implicit Differentiation
3.8 Related Rates
3.9 Linearization and Differentials

2. Limits and Continuity
2.1 Rates of Change and Tangents to Curves
2.2 Limit of a Function and Limit Laws
2.3 The Precise Definition of a Limit
2.4 One-Sided Limits
2.5 Continuity
2.6 Limits Involving Infinity; Asymptotes of Graphs

3. Differentiation
3.1 Tangents and the Derivative at a Point
3.2 The Derivative as a Function
3.3 Differentiation Rules
3.4 The Derivative as a Rate of Change
3.5 Derivatives of Trigonometric Functions
3.6 The Chain Rule
3.7 Implicit Differentiation
3.8 Related Rates
3.9 Linearization and Differentials



(c)

4.5 Monotonic Functions and the First Derivative Test

$$(\hat{\mathbf{W}}_{S_2}) = \arg\min_{\mathbf{W}_{S_2}} \sum_{i=1}^{2n_0-1} \lambda_i \left\| G_i^T(x, \mathbf{W}_T) - G_i^S(x, \mathbf{W}_{S_2}) \right\|_2^2, \quad (2)$$

Proposed Model Sequential Knowledge Transfer

- Step 3

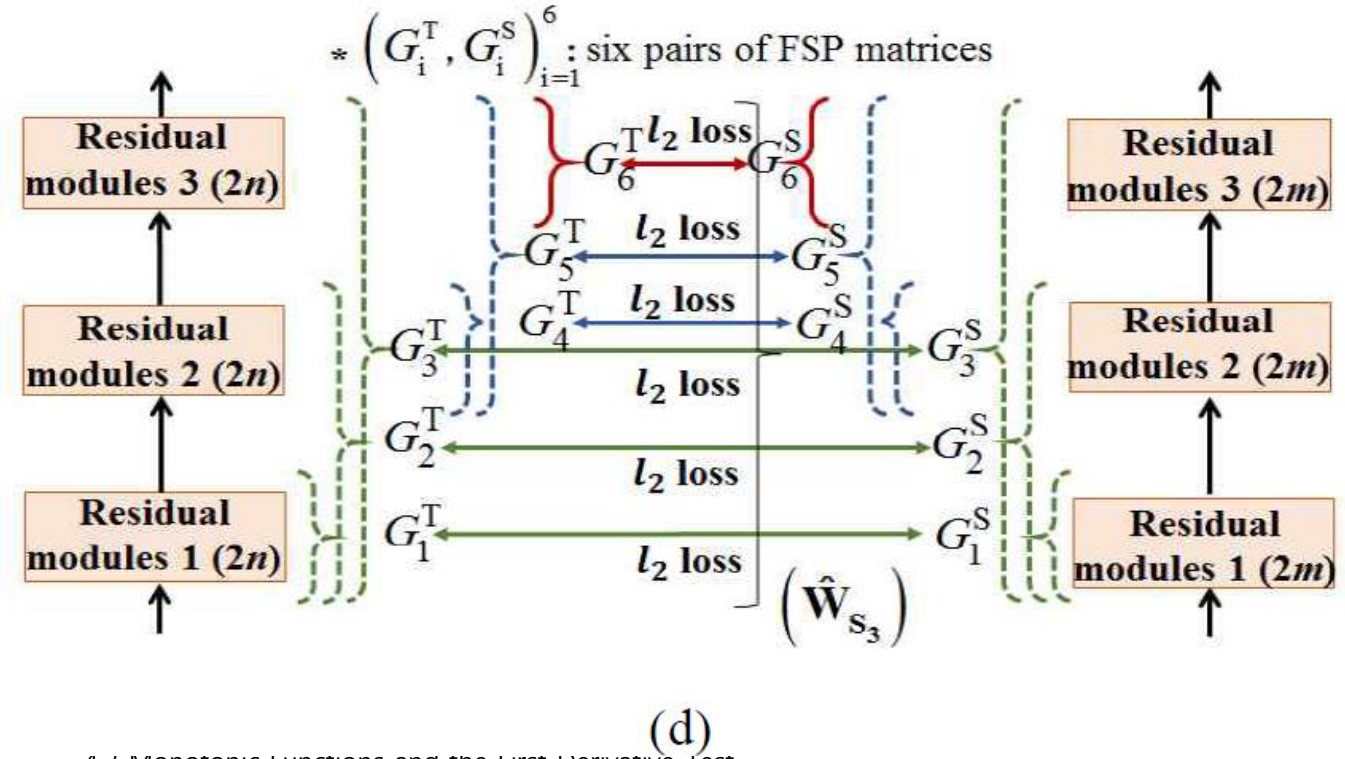
2. Limits and Continuity
2.1 Rates of Change and Tangents to Curves
2.2 Limit of a Function and Limit Laws
2.3 The Precise Definition of a Limit
2.4 One-Sided Limits
2.5 Continuity
2.6 Limits Involving Infinity; Asymptotes of Graphs

3. Differentiation
3.1 Tangents and the Derivative at a Point
3.2 The Derivative as a Function
3.3 Differentiation Rules
3.4 The Derivative as a Rate of Change
3.5 Derivatives of Trigonometric Functions
3.6 The Chain Rule
3.7 Implicit Differentiation
3.8 Related Rates
3.9 Linearization and Differentials

4. Applications of Derivatives
4.1 Extreme Values of Functions
4.2 The Mean Value Theorem
4.3 Monotonic Functions and the First Derivative Test
4.4 Concavity and Curve Sketching
4.5 Applied Optimization
4.6 Newton's Method
4.7 Antiderivatives

Proposed Model Sequential Knowledge Transfer

- Step 3



(d)

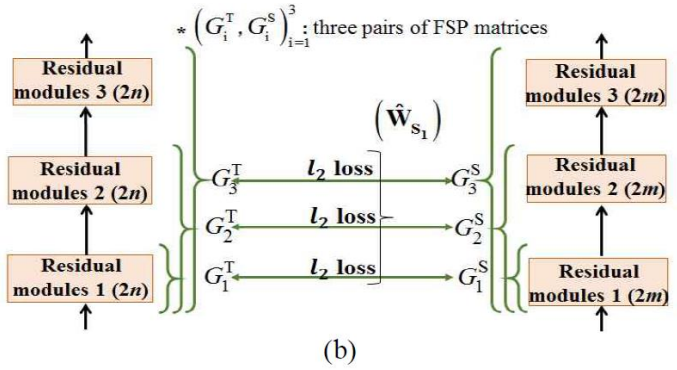
2. Limits and Continuity
2.1 Rates of Change and Tangents to Curves
2.2 Limit of a Function and Limit Laws
2.3 The Precise Definition of a Limit
2.4 One-Sided Limits
2.5 Continuity
2.6 Limits Involving Infinity; Asymptotes of Graphs

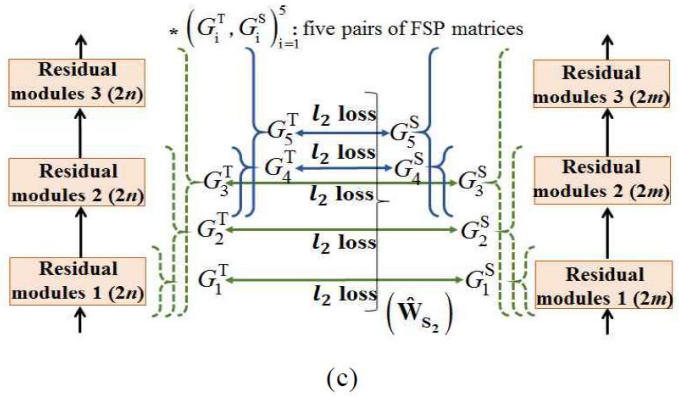
3. Differentiation
3.1 Tangents and the Derivative at a Point
3.2 The Derivative as a Function
3.3 Differentiation Rules
3.4 The Derivative as a Rate of Change
3.5 Derivatives of Trigonometric Functions
3.6 The Chain Rule
3.7 Implicit Differentiation
3.8 Related Rates
3.9 Linearization and Differentials

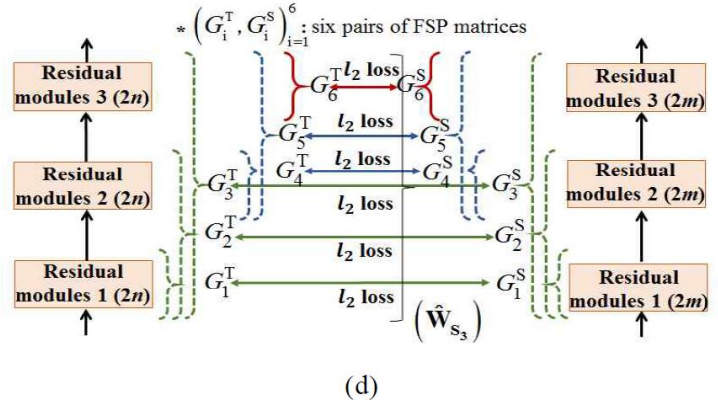
4.3 Monotonic Functions and the First Derivative Test
4.4 Concavity and Curve Sketching
4.5 Applied Optimization
4.6 Newton's Method
4.7 Antiderivatives

Proposed Model Sequential Knowledge Transfer

- Whole procedure







Step 1 Step 2 Step 3

Experimental Result

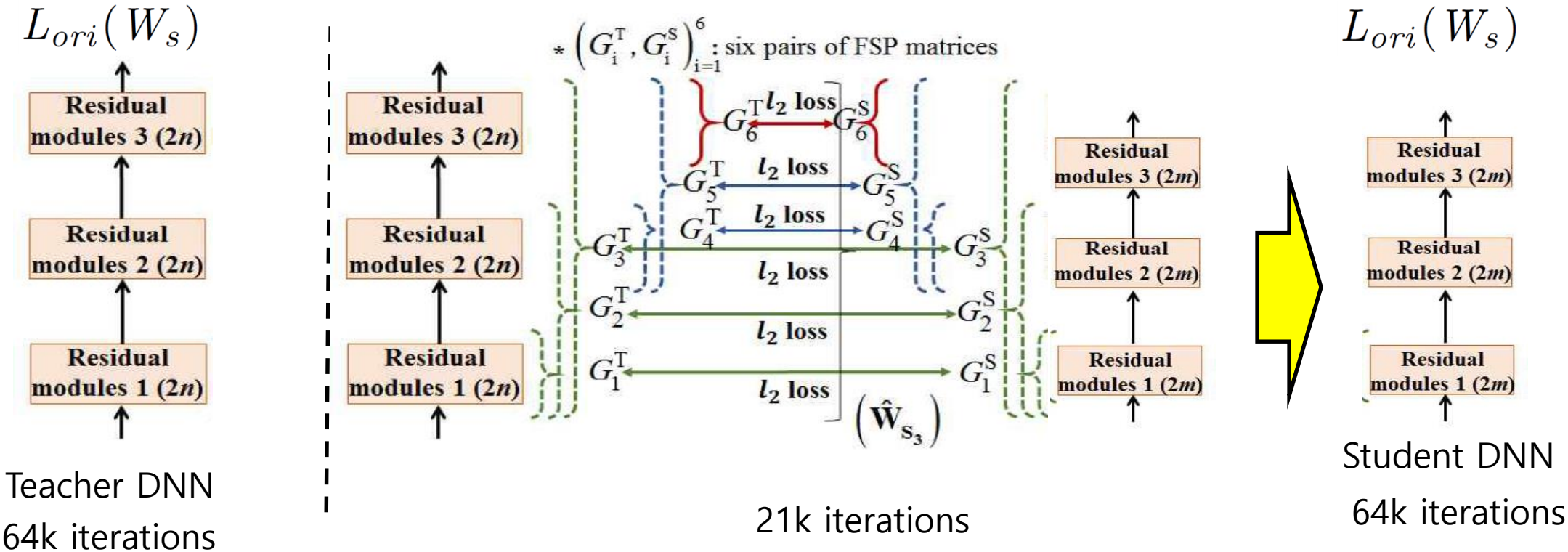
Experimental Result

- Contents
  - Performance Improvement
  - Network Minimization

Experimental Result

Performance improvement

- Experimental setting
  - CIFAR-10 dataset
  - Using 26-layers Residual network for the Teacher DNN
  - Using 8-layers Residual network for the Student DNN



Experimental Result

Performance improvement

- Experimental setting
  - CIFAR-10, Using 8-layers Residual network
  - 21k Iter for stage 1 and 64k Iter for stage 2
  - 26-layer teacher ResNet with an accuracy of 91.91%

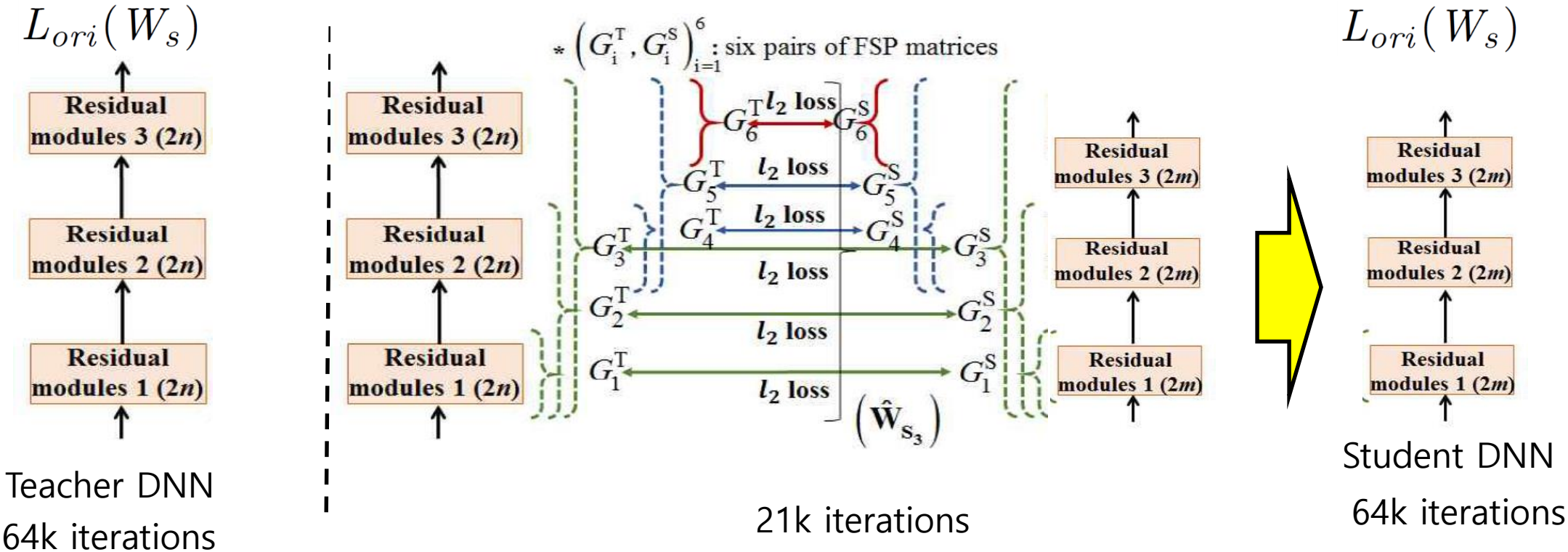
Method	Accuracy [%]	Reference
Hint-based method [25]	88.4	$P_c = 87.94\%$ for the original eight-layer ResNet
Original flow-based method [27]	88.72	
Proposed method	**88.96**	

CIFAR-10

Experimental Result

Performance improvement

- Experimental setting
  - CIFAR-100 dataset
  - Using 32-layers Residual network for the Teacher DNN
  - Using 14-layers Residual network for the Student DNN



Experimental Result

Performance improvement

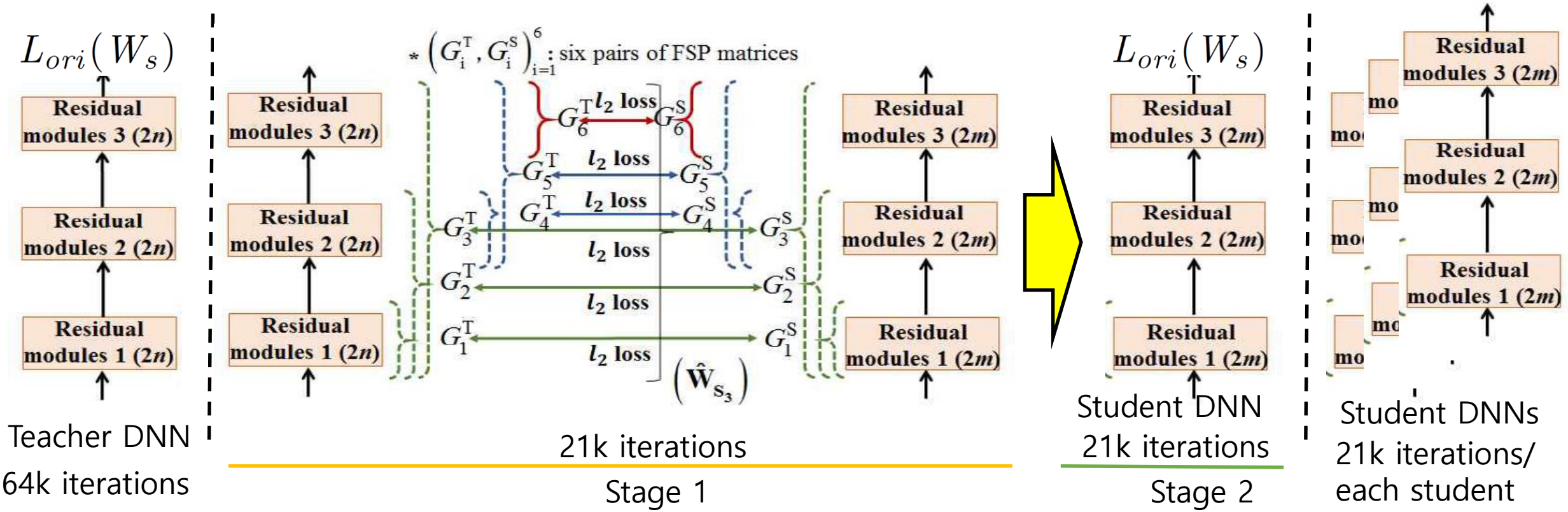
- Experimental setting
  - CIFAR-100, Using 14-layers Residual network
  - 32k Iter for stage 1 and 64k Iter for stage 2
  - 32-layer teacher ResNet with an accuracy of 64.69%

Method	Accuracy [%]	Reference
Hint-based method [25]	63.38	$P_c = 62.37\%$ for the original 14-layer ResNet
Original flow-based method [27]	64.74	
Proposed method	**65.06**	

CIFAR-100

Experimental Result Fast optimization

- Experimental setting
  - CIFAR-10 dataset
  - Using 26-layers Residual network for the Teacher and Student DNN



Experimental Result Fast optimization

- Experimental setting
  - CIFAR-10,
  - 21k Iter for stage 1, 21k Iter for stage 2

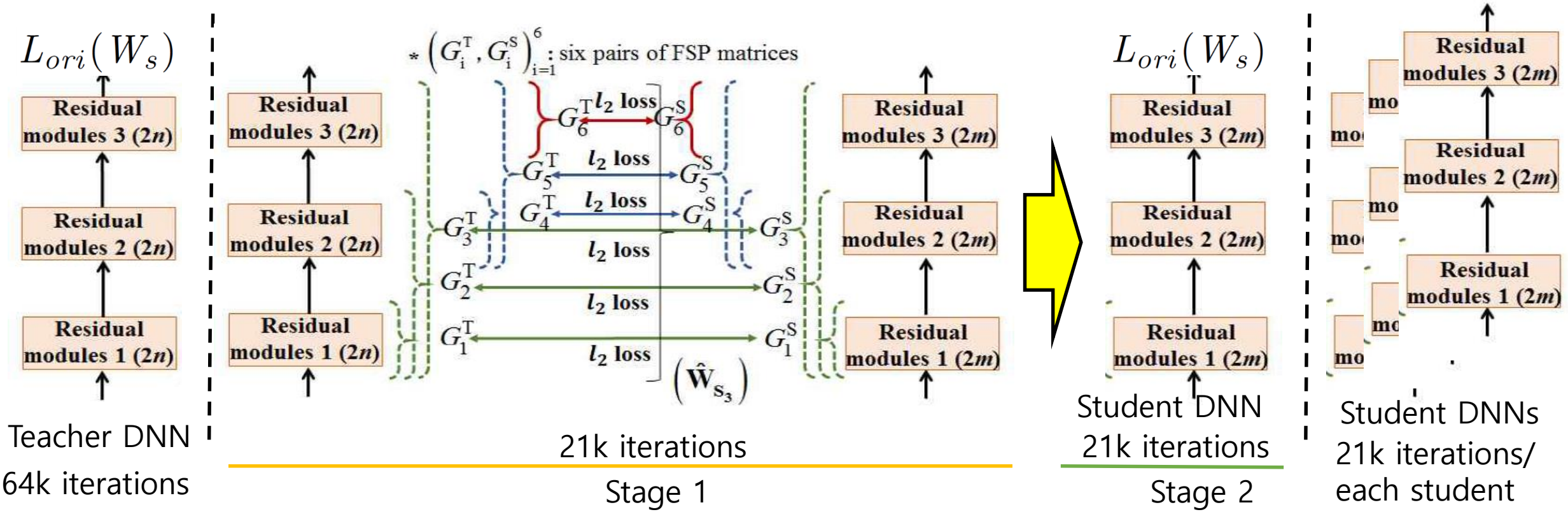
Method	Net1	Net2	Net3	Avg.	Ensemble	#Iter
Original teacher[a]	91.91	91.68	91.68	91.75	93.29	192k
Original teacher[b]	90.92	90.85	90.69	90.82	92.7	63k
Hint-based method [25]	92.07	91.75	91.81	91.87	93.02	138k
Original flow-based method [27]	91.84	92.13	92.25	92.07	93.59	126k
Proposed method	92.36	92.34	92.15	**92.28**	**93.68**	126k

a. *The 26-layer teacher ResNet was trained with 64,000 iterations.*

b. *The 26-layer teacher ResNet was trained with 21,000 iterations.*

Experimental Result Fast optimization

- Experimental setting
  - CIFAR-100 dataset
  - Using 32-layers Residual network for the Teacher and Student DNN



Experimental Result Fast optimization

- Experimental setting
  - CIFAR-100,
  - 32k Iter for stage 1, 21k Iter for stage 2

Method	Net1	Net2	Net3	Avg.	Ensemble	#Iter
Original teacher[a]	64.69	63.29	64.52	64.16	69.79	192k
Original teacher[b]	62.96	62.69	60.82	62.15	67.91	63k
Hint-based method [25]	63.54	64.43	64.07	64.01	68.68	168k
Original flow-based method [27]	64.16	64.3	64.48	64.31	69.5	159k
Proposed method	66.65	66.52	64.54	**65.9**	**69.98**	159k

a. *The 26-layer teacher ResNet was trained with 64,000 iterations.*

b. *The 26-layer teacher ResNet was trained with 21,000 iterations.*

Conclusion

- Propose a novel approach for **enhancing** knowledge distillation and knowledge transfer between teacher and student DNN models
- Help to obtain a **fast optimization** with high accuracy using the **densely distilled** flow-based knowledge and its **sequential** transfer
- Proposed method outperforms state-of-the-art knowledge transfer method in the **network minimization**