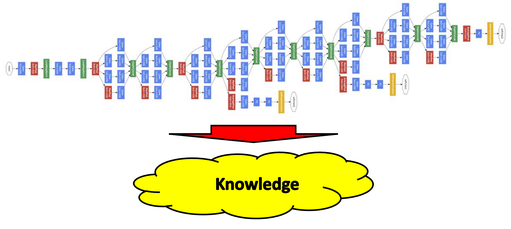
A Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning

Junho Yim [1] Donggyu Joo [1] Jihoon Bae [2] Junmo Kim [1]
{junho.yim, jdg105, junmo.kim}@kaist.ac.kr {baejh}@etri.re.kr
School of Electrical Engineering, KAIST, South Korea [1] Electronics and Telecommunications Research Institute [2]

IEEE 2017 Conference on Computer Vision and Pattern Recognition
CVPR July 21-26 2017 HONOLULU

Main Target : Knowledge Distillation

- Transfer the knowledge from the pre-trained DNN model to the new DNN model.

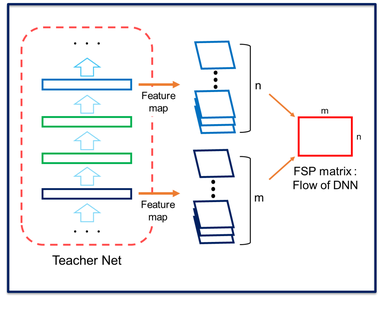


Proposed Distilled Knowledge

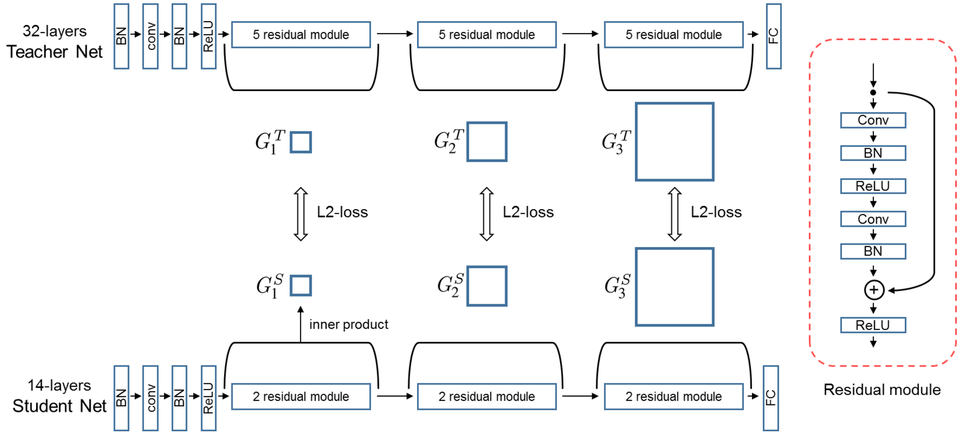
- Flow of solving a problem => Flow between two layers
- Computing the inner product between features from two layers
- Defined Knowledge : the FSP matrix

$$G_{i,j}(x;W) = \sum_{s=1}^{h}\sum_{t=1}^{w} \frac{F^1_{s,t,i}(x;W) \times F^2_{s,t,j}(x;W)}{h \times w}$$

$$x = \frac{-3 \pm \sqrt{17}}{4}$$
$$x = \frac{-3 \pm \sqrt{9+8}}{4}$$
$$x = \frac{-3 \pm \sqrt{(3)^2 - 4 \times 2 \times (-1)}}{2 \times 2}$$
$$2x^2 + 3x - 1 = 0$$
$$2x^2 + 3x = 1$$
$$Q : x = \frac{1}{2x+3}$$



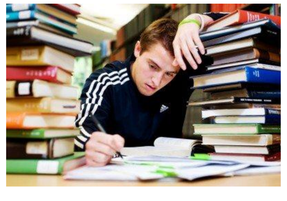
Proposed Architecture



- There are three points in the residual network for the CIFAR-10, 100 dataset where the spatial size changes. We selected several points to generate the FSP matrix.
- We minimize the distance between the FSP matrix of the student network and the one of the teacher network. The student network that went through the first stage is now trained by the main task loss at the second stage.

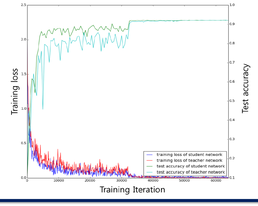
A Gift from Knowledge Distillation

- Fast Optimization
- Network Minimization
- Transfer Learning



Experiments – Fast Optimization

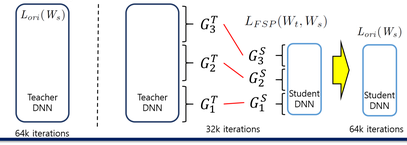
- By learning the flow of the solution procedure, the student network can study a task faster than usual
- Using 26-layers Residual network for the both Teacher and Student DNN



	Net 1	Net 2	Net 3	*Avg*	*Ensemble*	*#Iter*
Teacher	91.61	91.56	92.09	91.75	93.48	192k
Teacher *	90.47	90.83	90.62	90.64	92.6	63k
FitNet [20]*	91.69	91.85	91.64	91.72	92.98	98k
Student *	92.28	92.08	92.07	92.14	93.26	84k
Student *†	92.28	91.89	92.08	92.08	93.67	126k

Experiments – Network Minimization

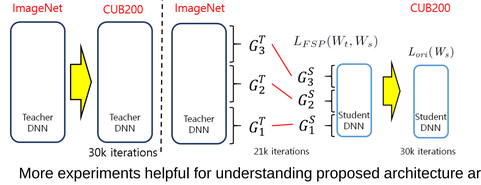
- Proposed method can improve the performance of a small student network by learning the distilled knowledge of a deep teacher network
- Using 32-layers Residual network for the Teacher DNN
- Using 14-layers Residual network for the Student DNN



	Accuracy
Teacher-original	64.06
Student-original	58.65
FitNet [20]	61.28
Proposed Method	63.33

Experiments – Transfer Learning

- The teacher DNN and student DNN can learn not only the same task, but also different tasks
- 34 layers Teacher DNN trained with ImageNet dataset
- 20 layers Student DNN fine-tuned with CUB200 dataset



	Accuracy
Teacher - fine tuning	77.72
Teacher - training from scratch	47.53
Student - training from scratch	47.73
FITNET [20]	70.19
Proposed Method	74.26

- More experiments helpful for understanding proposed architecture are stated in the main paper.