ECCV 2018

Self-supervised Knowledge Distillation Using Singular Value Decomposition

Seung Hyun Lee*, Dae Ha Kim, Byung Cheol Song

{lsh910703, kdhht5022}@gmail.com, bcsong@inha.ac.kr

Department of Electronic Engineering, Inha University, Republic of Korea

CVIP Computer Vision & Image Processing.

Introduction

Knowledge Distillation

- Enhance shallow and simple network by transferring deep and complex network's knowledge.

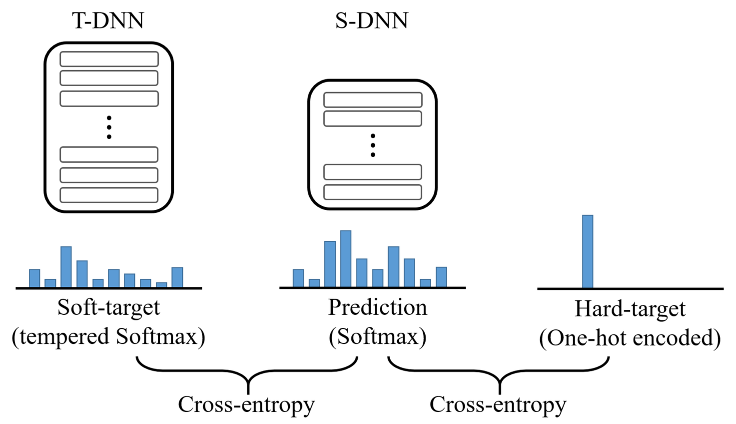
Contribution Points

- Define novel knowledge having rich information.
- Overcome limitation(s) of conventional transfer learning architecture.
- Additionally enhance the performance through multi-task learning.

Related works

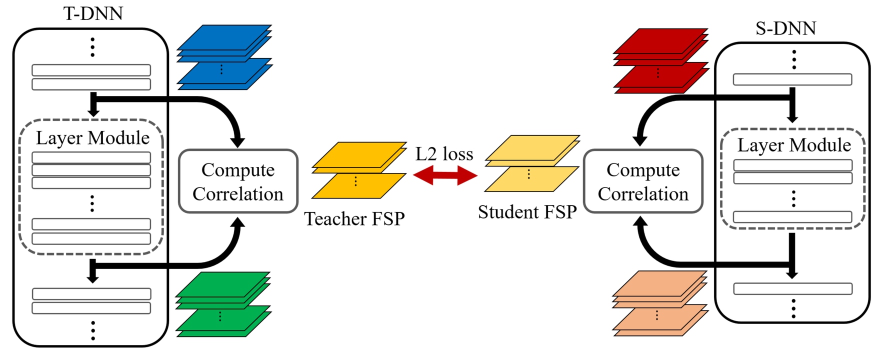
Soft-target[1]

- Define knowledge as softened teacher's prediction
- Pros
  - Easy to build
  - Multi-task learning
  - Not relevant to structure
- Cons
  - Available to the same domain
  - Too naïve knowledge



Flow of Solving Problem (FSP) [2]

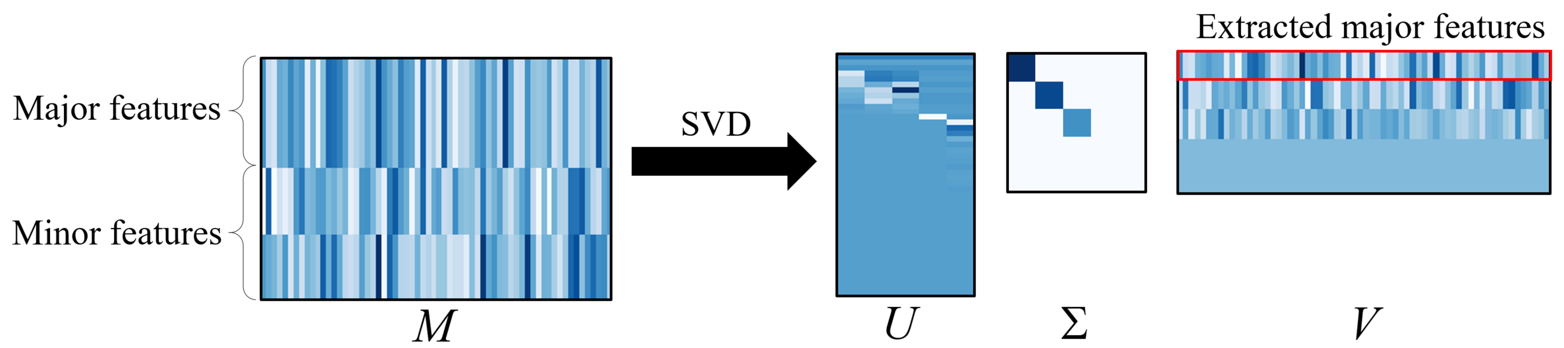
- Define knowledge as feature correlation so-called 'Flow of Solving Problem.'
- Compute correlation by cross-product and compress it by averaging
- Pros
  - Rich knowledge
  - Not relevant to domain
- Cons
  - Available to similar structure
  - Too naïve knowledge
  - 2-stage learning



Singular Value Decomposition (SVD)

- Decompose the matrix into singular vectors and singular values.
- Decomposed singular vectors contain compressed information while maintaining major features.

→ **Effectively compress feature map while maintaining important features.**



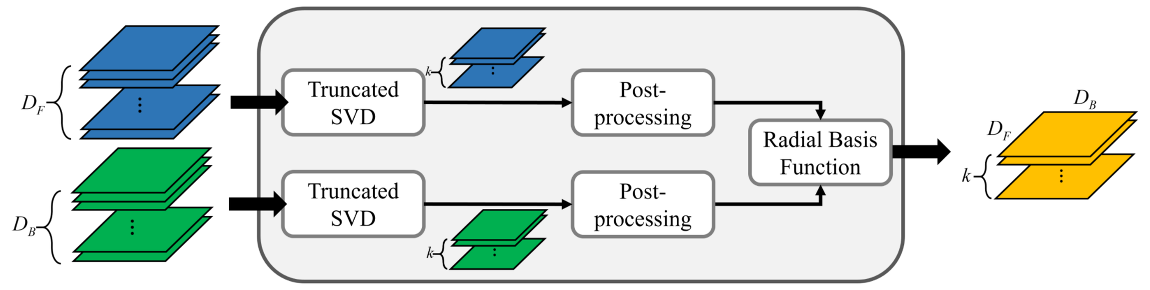
- Gradient of SVD is defined in previous work[3].

$$\frac{\partial L \circ f}{\partial X} = DV^T + U\left(\frac{\partial L}{\partial \Sigma} - U^T D\right)_{diag} V^T + U\Sigma\left(K^T \circ \left(V^T\left(\frac{\partial L}{\partial V} - VD^T U\Sigma\right)\right)\right)_{sym} V^T$$

Method

Distillation Module

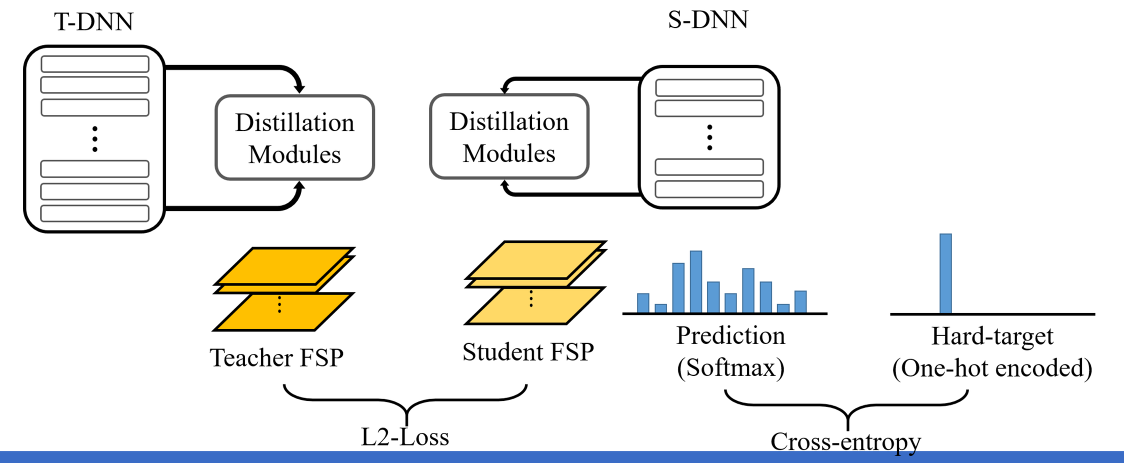
- Distill rich knowledge by SVD and RBF.
- Overcome the limitations of normal transfer learning networks.



Multi-task Learning

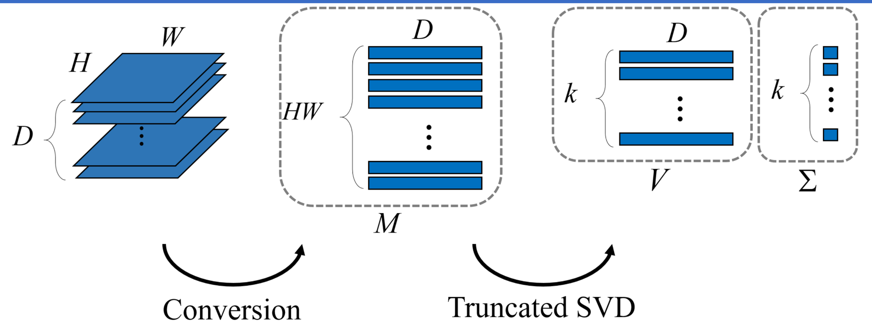
- Simultaneous learning of main-task and self-supervised task (knowledge distillation) improves main-task performance.
- Employ gradient clipping for learning focused on main-task.

$$\nabla(\theta_S)^{clipped}_{transfer} = \begin{cases} \frac{1}{1+\exp(-\tau + p)} \nabla(\theta_S)_{transfer} & \nabla(\theta_S)_{transfer} < \nabla(\theta_S)_{main} \\ \nabla(\theta_S)_{transfer} & \text{otherwise} \end{cases}, \quad \tau = \frac{\left\|\nabla(\theta_S)_{main}\right\|_2}{\left\|\nabla(\theta_S)_{transfer}\right\|_2}$$



Feature Compression using Truncated SVD

- Increase the degree of freedom to calculate the correlation via SVD.
- Compress and maintain rich information by adopting SVD.
- Rearrange gradient to reduce unnecessary costs.

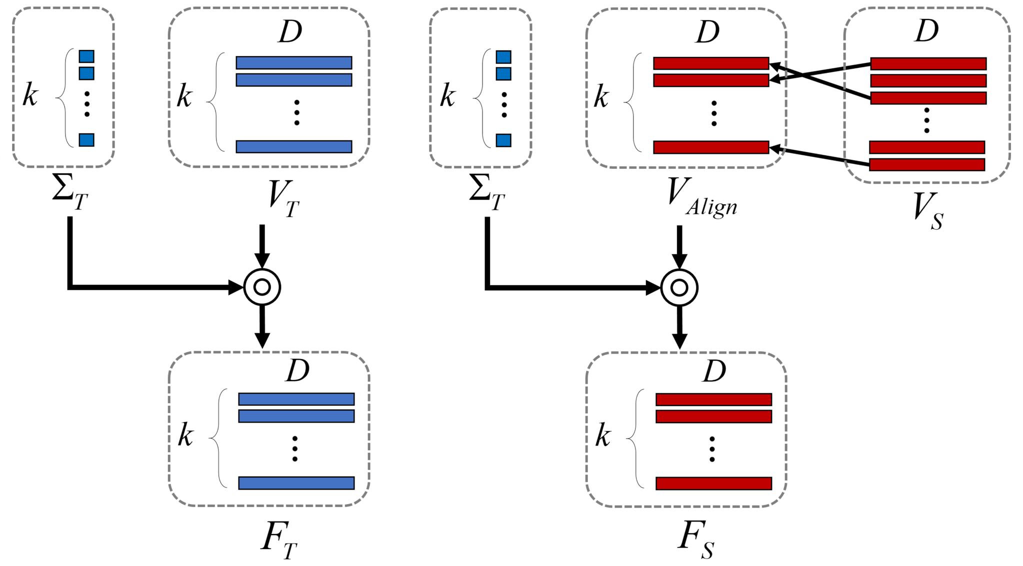


$$\nabla(M) = \begin{cases} UE^T - U(E^T V)_{diag} V^T - 2U(K \circ (\Sigma^T V^T E))sym\Sigma^T V^T & HW \le D \\ 2U\Sigma(K^T \circ (V^T \nabla(V)))_{sym} V^T, & \text{otherwise} \end{cases}$$

$$E = \nabla(V)\Sigma^{-1}, \quad K = \begin{cases} \frac{1}{\sigma_i^2 - \sigma_j^2} & i \ne j, (1 \le i, j \le k) \\ 0 & \text{otherwise} \end{cases}$$

Post-processing

- (1) Singular vectors which have similar singular values may be decomposed in random order. (2) Singular vectors may be decomposed into opposite directions.
- Align the order and direction according to cosine similarity with the teacher singular vector.



$$F_S = \left\{ \frac{\sigma_{T,i}}{\left\|\Sigma_T\right\|_2} \mathbf{v}_{Align,i} \right\} \qquad s_j = \arg\max_j \left(\left| \mathbf{v}_{T,i} \cdot \mathbf{v}_{S,j} \right| \right), (1 \le i \le k), (1 \le j \le k+1)$$

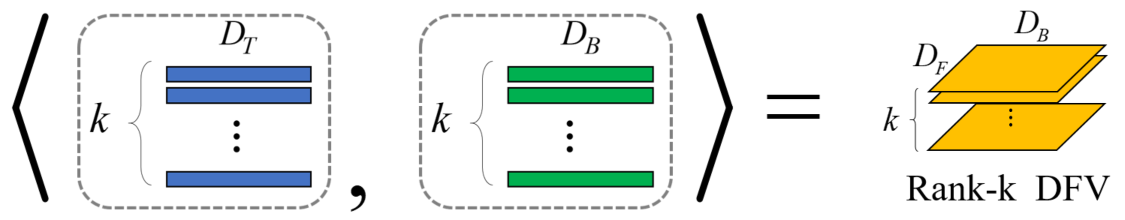
$$\mathbf{v}_{Align,i} = \mathbf{v}_{S,s_j}$$

Radial Basis Function

- Maintain gradient flow by using RBF which contain exponential function.

$$DFV = \left\{ \exp\left(-\frac{\left\| f^{FFM}_{m,l} - f^{BFM}_{n,l} \right\|_2^2}{\beta}\right), 1 \le m \le D_F, 1 \le n \le D_B, 1 \le l \le k \right\}$$

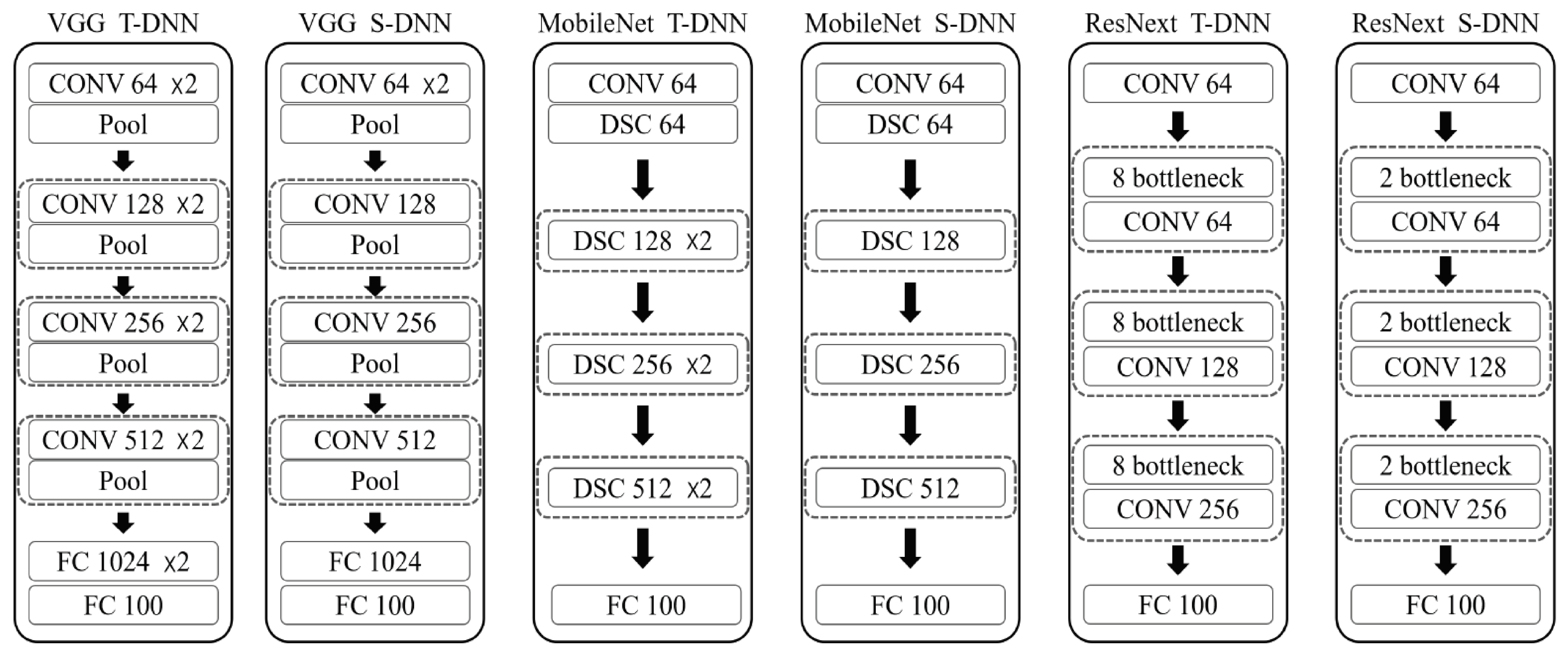
$$f_{T,i} = \frac{\sigma_{T,i}}{\left\|\Sigma_T\right\|_2} \mathbf{v}_{T,i} \qquad F_T = \left\{ f_{T,i} \mid 1 \le i \le k \right\}$$



Experimental results

Verification on Small-size Dataset (CIFAR100)

- VGG, MobileNet : Network structures well normalized with fewer parameters, so they respond sensitively to additional information.
- ResNext : Network structures well regularized with fewer parameters.



Small Network Enhancement

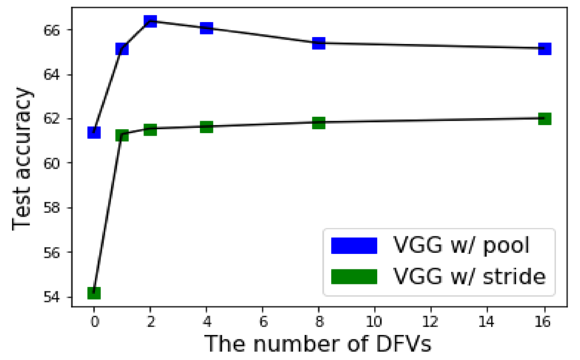
- Dramatically enhanced in VGG and VGG_stride.
- Less enhanced in ResNext because of well-regularized architecture.

Network	Model	FLOPs	Params	Accuracy
VGG	T-DNN	576.3M	10.9M	64.44
	S-DNN	121.3M	3.8M	61.37
	FSP	121.3M	3.8M	64.54
	proposed	121.3M	3.8M	**65.05**
MobileNet	T-DNN	98.4M	2.3M	57.85
	S-DNN	37.8M	0.82M	56.15
	FSP	37.8M	0.82M	56.53
	proposed	37.8M	0.82M	**58.15**
ResNext	T-DNN	547.3M	0.66M	66.58
	S-DNN	247.6M	0.34M	64.00
	FSP	247.6M	0.34M	63.60
	proposed	247.6M	0.34M	**65.43**
VGG_stride	T-DNN	576.3M	10.9M	64.44
	S-DNN	15.6M	3.8M	54.17
	Proposed	15.6M	3.8M	**61.15**

Multi-task Learning

Model	Mechanism	Accuracy
FSP	2 Stage	64.54
	1 stage	64.89
Proposed	2 Stage	65.05
	1 stage	**65.54**

The Number of DFVs



Verification on Large-size Dataset (Tiny-imagenet and Imagenet-50)

Data set	Model	Accuracy	Data set	Model	Accuracy
Tiny-imagenet	T-DNN	53.37	Imagenet-50	T-DNN	70.09
	S-DNN	47.32		S-DNN	65.03
	FSP	48.77		FSP	69.59
	Proposed	**51.82**		Proposed	**71.04**

[1] G. Hinton et al., "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*(2015)
[2] J. Yim et al., "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," CVPR 2017
[3] C. Ionescu et al., "Training deep networks with structured layers by matrix backpropagation," ICCV 2015