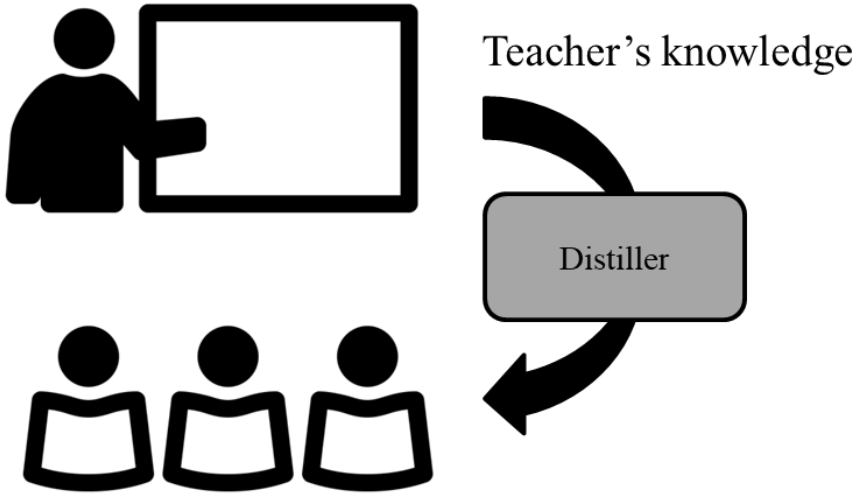
Knowledge Distillation Overview & Our Research Results

Knowledge Distillation Overview

- What is knowledge distillation?
  - 한 network의 information을 다른 network에게 전달하여 성능을 개선하는 알고리즘

Teacher **Student**

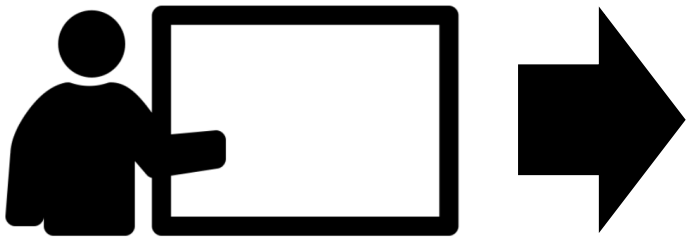


Knowledge Distillation Overview

- What is knowledge distillation?
  - 목적에 맞도록 주입할 knowledge를 정의
    ex) Better inference results, stable guidance

  - Network의 knowledge을 distiller를 통해 transferrable form으로 변환
    - Teacher network의 constraints를 조절
    - Student가 학습할 information을 추출

  → Transfer할 **knowledge**와 **distiller**을 결정하는 것이 핵심



Teacher's knowledge

Smaller network : Small network enhancement

Different domain : Domain adaptation

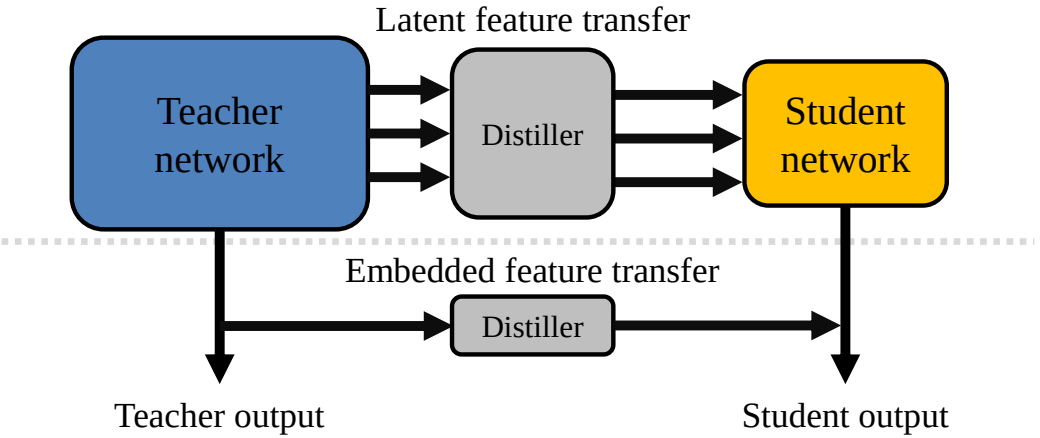
Unlabeled dataset : Semi-supervised learning

Knowledge Distillation Overview

- Knowledge distillation의 종류
  - Transfer할 feature에 따른 분류

  - Embedded feature transfer
    - 처음으로 제안된 knowledge distillation 기법 [1]
    - 높은 task dependency
    - 상대적으로 information의 양이 적음

  - Latent feature transfer
    - Feature maps을 전달하여 더 명확한 guidance를 제공
    - 높은 architecture dependency
    - 더 많은 사용자의 지식이 더 많이 요구됨 [2]

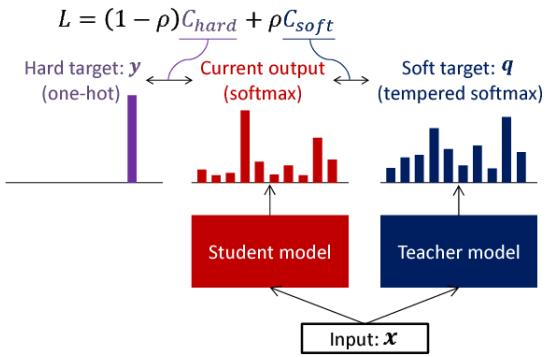


[1] Geoffrey Hinton, et al. Distilling the knowledge in a neural network. arXiv:1503.02531, 2015.
[2] Heo, Byeongho, et al. A comprehensive overhaul of feature distillation. ICCV 2019

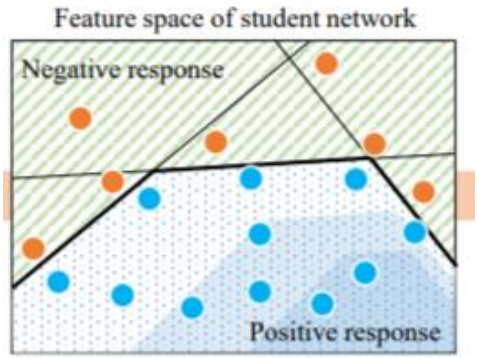
Knowledge Distillation Overview

- Knowledge distillation의 종류
  - Distiller를 설계하는 방식에 따른 분류

  - Smoothing function
    - Feature를 smoothen하여 over-constraints를 방지
      ex) Soft target [1], activation boundary[3]



Soft target [1]



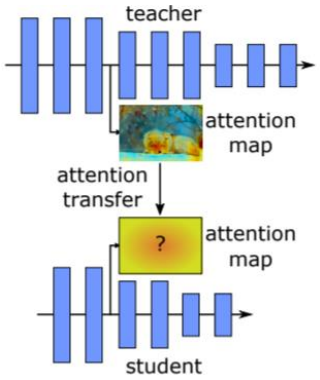
Activation boundary [3]

[3] Byeongho Heo, et. al. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. AAAI2019

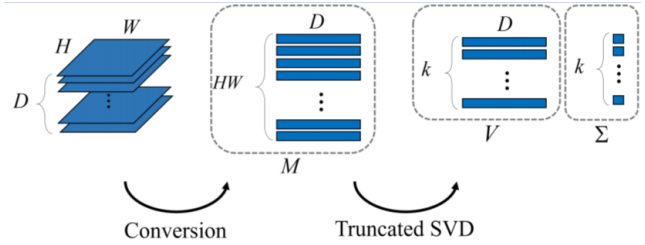
Knowledge Distillation Overview

- Knowledge distillation의 종류
  - Distiller를 설계하는 방식에 따른 분류

  - Extract information
    - Feature에서 특정 information만 추출
      **Feature의 형태 자체가 변경
      ex) attention map [4], singular vectors [5]



Attention map [4]



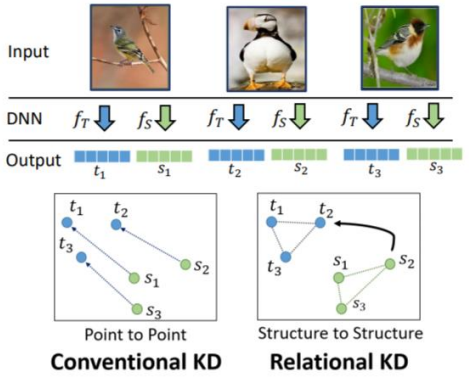
Singular vectors [5]

[4] Zagoruyko, Sergey et. al. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv: 2016.
[5] Seunghyun Lee, et. al. Self-supervised knowledge distillation using singular value decomposition. ECCV 2018

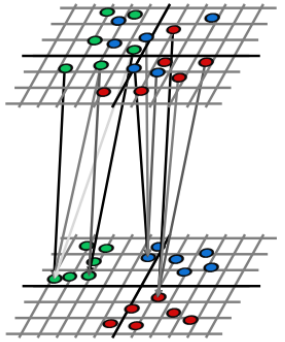
Knowledge Distillation Overview

- Knowledge distillation의 종류
  - Distiller를 설계하는 방식에 따른 분류

  - Composing a higher-level information
    - Feature를 조합하여 더 고차원의 정보로 변환
      ex) inter-data relation [6], embedding procedure [7]



Inter-data relation [6]

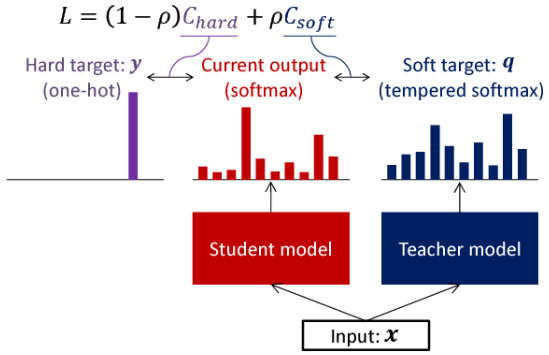


Embedding procedure [7]

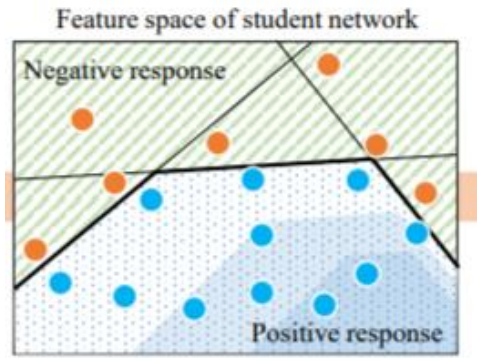
[6] Wonpyo Park, et. al. Relational Knowledge Distillation. CVPR2019
[7] Seunghyun Lee, Byung Cheol Song. Graph-based Knowledge Distillation by Multi-head Attention Network. BMVC2019

Knowledge Distillation Overview

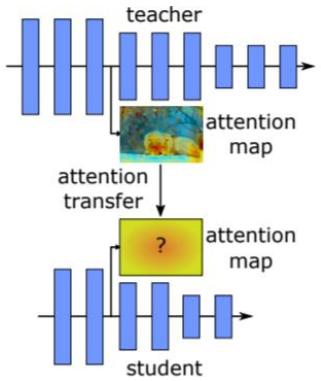
- Knowledge distillation의 주요 issue
  - There is no common agreed “knowledge”.
    → 이게 진짜 CNN의 knowledge일까?



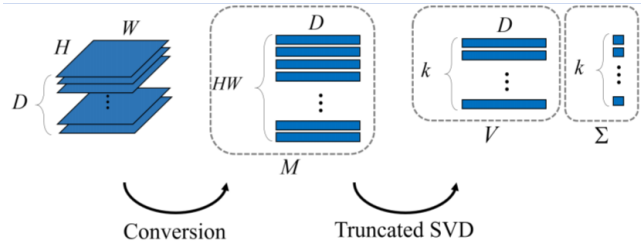
Soft target [1]



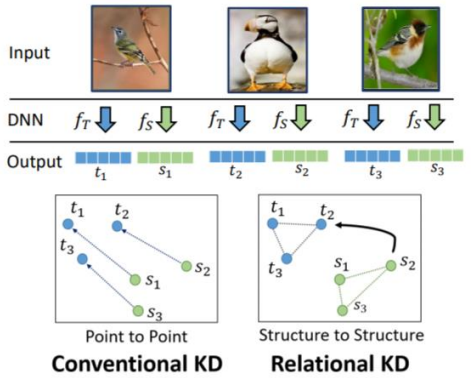
Activation boundary [3]



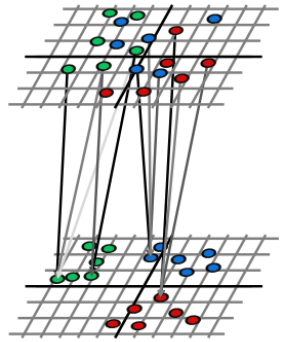
Attention map [4]



Singular vectors [5]



Inter-data relation [6]



Embedding procedure [7]

Knowledge Distillation Overview

- Knowledge distillation의 주요 issue
  - Require lots of human knowledge
    - feature map sensing 위치
    - 각종 hyper-parameter
    - knowledge 학습 방식
    - 등등…

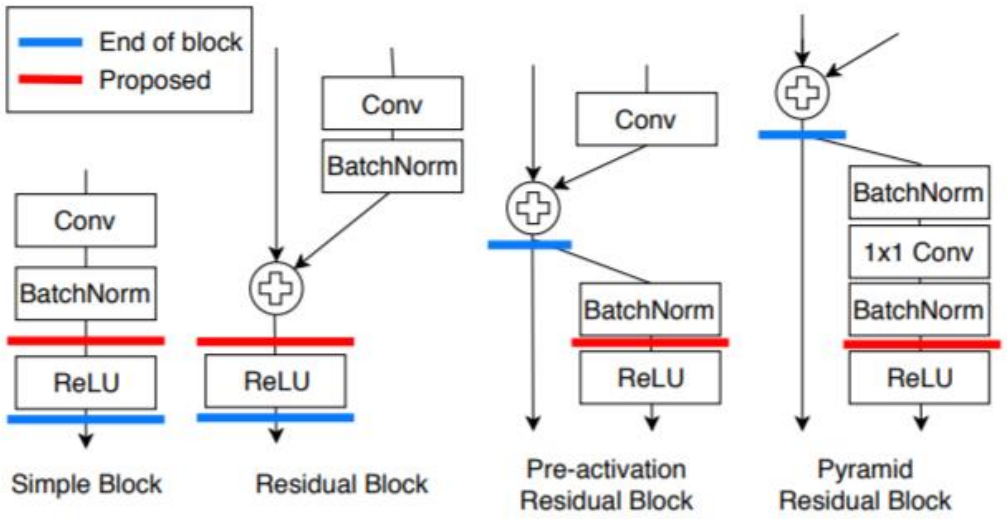


Figure from [8]

[8] Heo, Byeongho, et al. A comprehensive overhaul of feature distillation. ICCV 2019

Knowledge Distillation Overview

- Knowledge distillation의 주요 issue
  - How to choose a teacher network

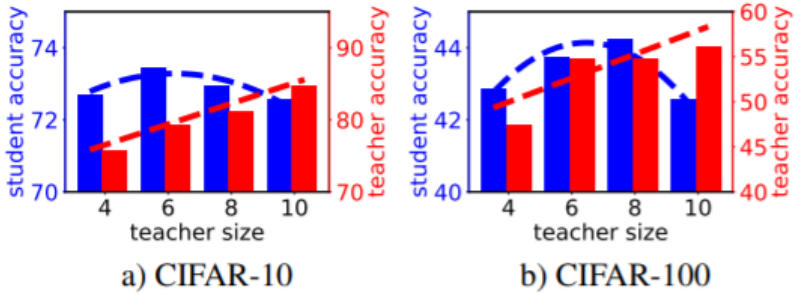


Figure from [9]

Network Types		CIFAR-10						CIFAR-100					
		Independent		DML		DML-Ind		Independent		DML		DML-Ind	
Net 1	Net 2	Net 1	Net 2	Net 1	Net 2	Net 1	Net 2	Net 1	Net 2	Net 1	Net 2	Net 1	Net 2
Resnet-32	Resnet-32	92.47	92.47	92.68	92.80	0.21	0.33	68.99	68.99	71.19	70.75	2.20	1.76
WRN-28-10	Resnet-32	95.01	92.47	95.75	93.18	0.74	0.71	78.69	68.99	78.96	70.73	0.27	1.74
MobileNet	Resnet-32	93.59	92.47	94.24	93.32	0.65	0.85	73.65	68.99	76.13	71.10	2.48	2.11
MobileNet	MobileNet	93.59	93.59	94.10	94.30	0.51	0.71	73.65	73.65	76.21	76.10	2.56	2.45
WRN-28-10	MobileNet	95.01	93.59	95.73	94.37	0.72	0.78	78.69	73.65	80.28	77.39	1.59	3.74
WRN-28-10	WRN-28-10	95.01	95.01	95.66	95.63	0.65	0.62	78.69	78.69	80.28	80.08	1.59	1.39

Table from [10]

Tag	Model name	Input size	Top-1 accuracy
T(A)	EfficientNet-B7 [35]	600	84.4
T(B)	PNASNet-large [18]	331	82.9 74
T(C)	SE-ResNet-154 [11]	224	81.33
T(D)	PolyNet [41]	331	81.23
T(E)	Inception-ResNet-v2 [32]	299	80.217
T(F)	ResNeXt-101 [38]	224	79.431
T(G)	Wide-ResNet-101 [40]	224	78.84
T(H)	ResNet-152 [5]	224	78.31

Teacher models				
GT	T(A)	T(B)	T(E)	T(F)
S_3 (65.6)	S_2 (66.6)	S_3 (66.9)	S_1 (67.4)	S_5 (67.1)
S_4 (65.6)	S_3 (66.5)	S_5 (66.4)	S_4 (67.0)	S_1 (67.1)
S_5 (65.5)	S_4 (66.3)	S_4 (66.1)	S_5 (66.9)	S_4 (66.6)
S_1 (65.5)	S_5 (66.0)	S_1 (65.7)	S_3 (66.5)	S_3 (66.3)
S_2 (65.4)	S_1 (65.8)	S_2 (65.4)	S_2 (66.1)	S_2 (66.0)

Tables from [11]

[9] Zhang, Ying, et al. "Deep mutual learning." CVPR. 2018.
[10] Mirzadeh, Seyed Iman, et al. "Improved knowledge distillation via teacher assistant." AAAI 2020.
[11] Liu, Yu, et al. "Search to distill: Pearls are everywhere but not the eyes." CVPR 2020.

Knowledge Distillation Overview

- Summary
  - Knowledge distillation은 오래 전부터 연구되어 왔지만 아직 blue ocean으로 생각됨

  - 이론적으로도 중요한 기법이지만 practical한 활용 및 성능 개선에 치우쳐져 있음
    → 너무 많은 hyper-parameters로 검증이 어려움..

Knowledge Distillation Overview

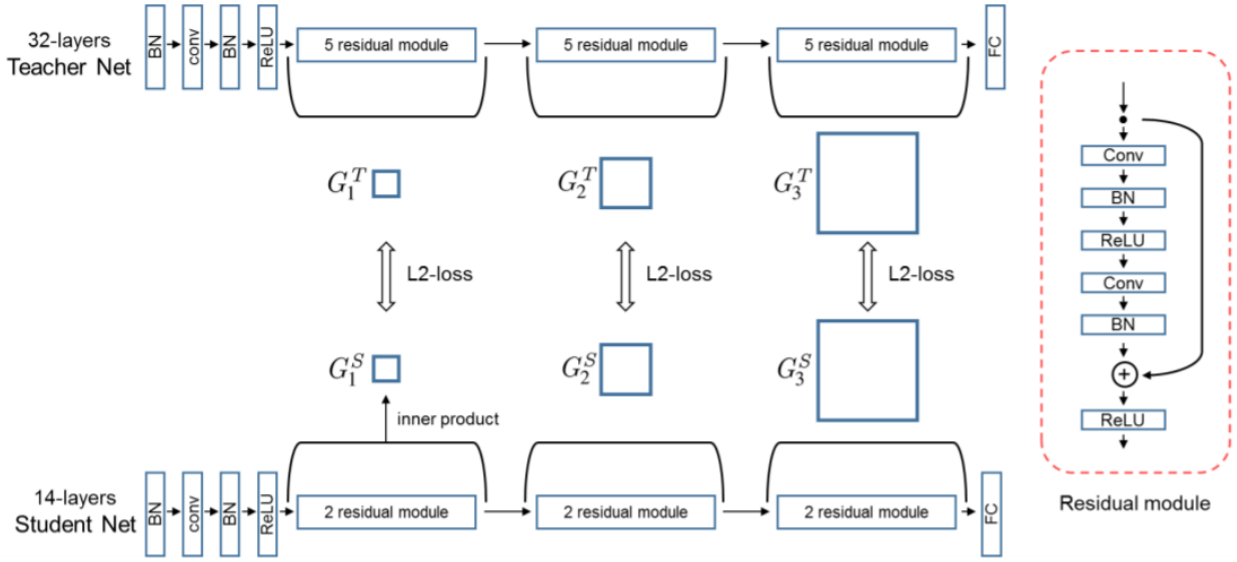
- Summary
  - Knowledge distillation은 오래 전부터 연구되어 왔지만 아직 blue ocean으로 생각됨
  - 이론적으로도 중요한 기법이지만 practical한 활용 및 성능 개선에 치우쳐져 있음
    → 너무 많은 hyper-parameters로 검증이 어려움..

** NN의 knowledge라고 납득이 되면서 성능 개선도 얻을 수 있는 방법이 있을까?

Our Research Results

- Insight: Flow of solving problem (FSP) [12]
  - Teacher network의 두 지점의 feature maps를 sensing
  - Gramian matrix를 통해 두 feature maps의 relation을 정의
    → style transfer와 유사
  - 계산된 relation이 feature map의 변화 과정, 즉, CNN의 FSP로 정의

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



[12] Yim, Junho, et al. "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning." CVPR 2017.

Our Research Results

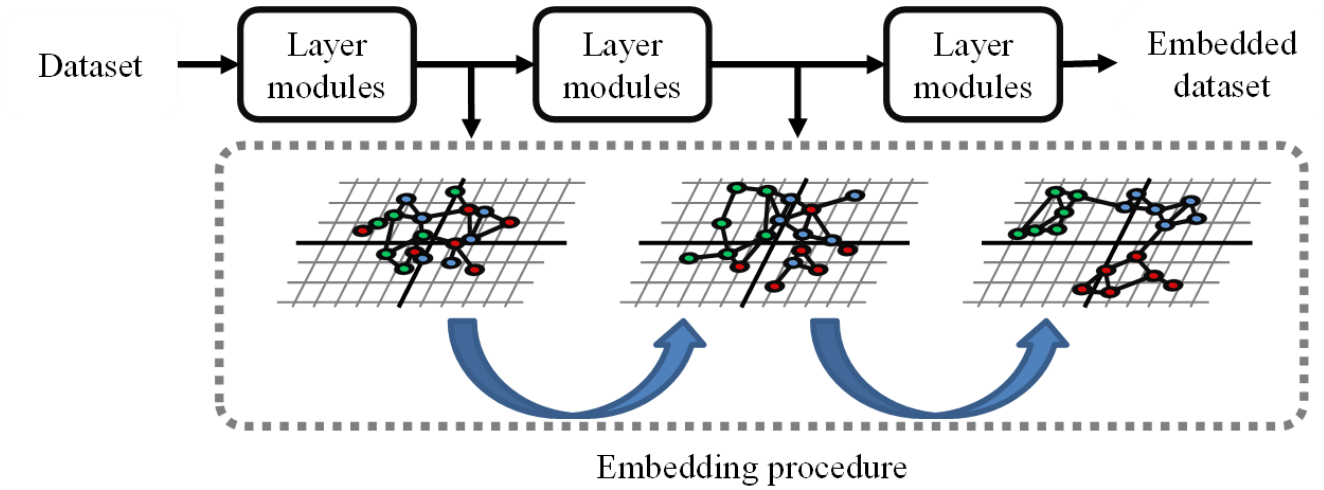
- Insight: Flow of solving problem (FSP) [12]
  - Teacher network의 두 지점의 feature maps를 sensing
  - Gramian matrix를 통해 두 feature maps의 relation을 정의
    → style transfer와 유사
  - 계산된 relation이 feature map의 변화 과정, 즉, CNN의 FSP로 정의
  - 의문점:
    - Feature map의 변화 과정이 FSP를 표현할 수 있을까?
      → CNN의 FSP는 무엇일까?
    - Gramian matrix가 feature maps의 relation을 표현하기 적절한 방법일까?

[12] Yim, Junho, et al. "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning." CVPR 2017.

Our Research Results

- CNN의 FSP는 무엇일까?
  - CNN을 사용하는 목적은 high-dimensional data를 low-dimensional space로 mapping하여 분석하기 쉽도록 만드는 것 → Embedding

  - Embedding을 위해 inter-data relation을 계산할 필요가 있음



Our Research Results

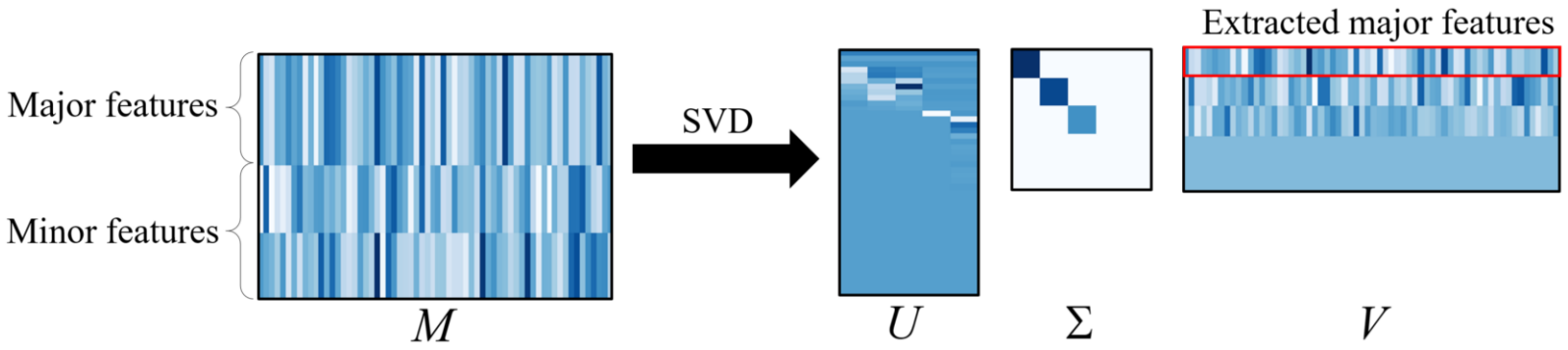
- Gramian matrix가 feature maps의 relation을 표현하기 적절한 방법일까?
  - Grammian matrix
    → Sum of Kronecker product
    - Kronecker product: 각 단계의 feature relation을 계산
    - sum: 계산된 relation의 크기를 압축

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

  - Inter-data relation을 표현하기 위해 두 function을 대체할 기법이 필요
    - Set of feature maps을 압축
    - Inter-data relation을 계산

Our Research Results

- Knowledge distillation using singular value decomposition [5, 13]
  - Feature map은 spatial relation이 높음
    → 몇 개의 major한 information을 가진 feature vectors로 표현가능
  - Feature map을 conversion하여 set of feature vectors로 표현 및 SVD로 decompose
  - Right-hand singular vectors를 압축된 feature vector로 정의
    → 더 적은 연산량으로 handling할 수 있음



[13] Lee, Seunghyun, and Byung Cheol Song. "Knowledge Transfer via Decomposing Essential Information in Convolutional Neural Networks." TNNLS (2020).

Our Research Results

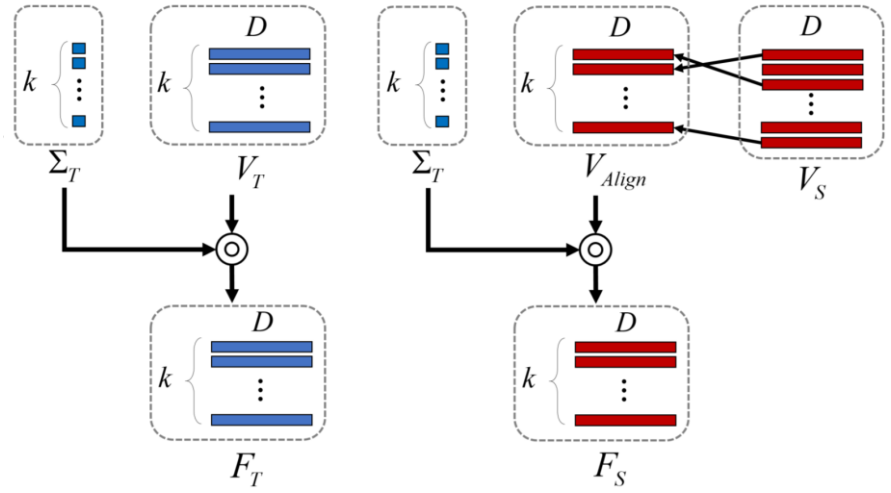
- Knowledge distillation using singular value decomposition
  - Singular vectors가 가진 두 문제를 해결하여 transferrable하게 변환
    → Sign ambiguity

$$USV = (-U)S(-V)$$

    → Order ambiguity

  - Student가 teacher와 어느 정도 유사할 것이라고 가정
    → similarity의 절대값이 가장 높은 것끼리 matching

```
def Align_rsv(x, y):
    cosine = tf.matmul(x, y, transpose_a=True)
    mask = tf.where(tf.equal(tf.reduce_max(tf.abs(cosine), 1,keepdims=True), tf.abs(cosine)),
                    tf.sign(cosine), tf.zeros_like(cosine))
    x = tf.matmul(x, mask)
    return x, y
```



Our Research Results

- Knowledge distillation using singular value decomposition
  - 실험 결과

Datasets	-	Student	SL	Fitnet	FSP	DML	AB	RKD	MHGD	SVD
CIFAR10	Last Acc.	89.24	90.03	88.95	88.91	89.93	91.06	90.72	90.83	90.91
	Best Acc.	89.37	90.03	89.18	89.01	90.16	91.19	90.79	91.22	90.92
	Iter. time	0.077	0.114	0.142	0.143	0.307	0.160	0.115	0.381	0.392
CIFAR100	Last Acc.	71.68	71.79	72.74	71.56	73.27	73.08	73.40	73.98	73.64
	Best Acc.	71.94	72.08	72.96	71.70	73.47	73.41	73.48	**74.30**	73.78
	Iter. time	0.077	0.114	0.142	0.143	0.307	0.160	0.115	0.381	0.392
TINY	Last Acc.	60.67	61.04	60.72	61.20	62.34	61.49	62.16	62.17	63.34
	Best Acc.	61.22	61.52	61.25	61.49	62.61	61.95	62.37	62.52	63.51
	Iter. time	0.177	0.263	0.326	0.332	0.694	0.369	0.260	0.602	0.631
CUB200	Last Acc.	38.89	51.73	54.42	42.29	52.49	62.60	30.96	62.54	62.37
	Best Acc.	39.02	51.97	54.78	42.60	52.50	62.60	31.24	62.70	62.48
	Iter. time	0.615	0.707	0.860	0.858	1.444	0.906	0.717	1.243	1.289

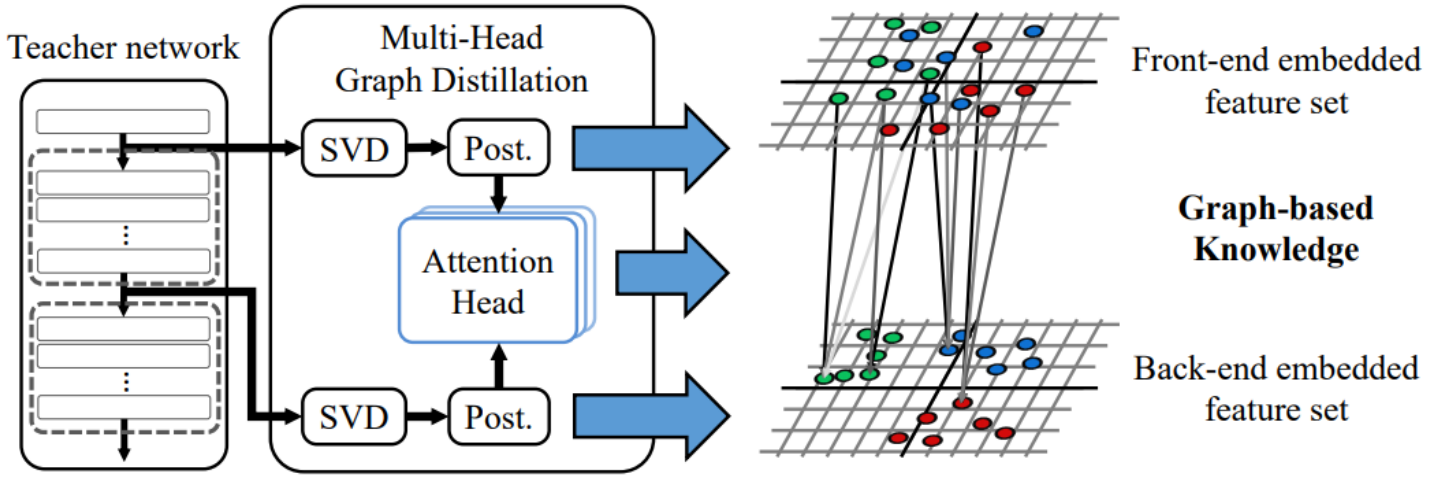
TABLE VIII

PERFORMANCE COMPARISON ACCORDING TO THE NUMBER OF SINGULAR VECTORS [%]. BOLD INDICATES THE BEST PERFORMANCE

N	1	2	4	8	16	$\hat{N}$ in Eq. (14)
Accuracy	73.57	73.76	73.84	73.86	73.71	**74.05**

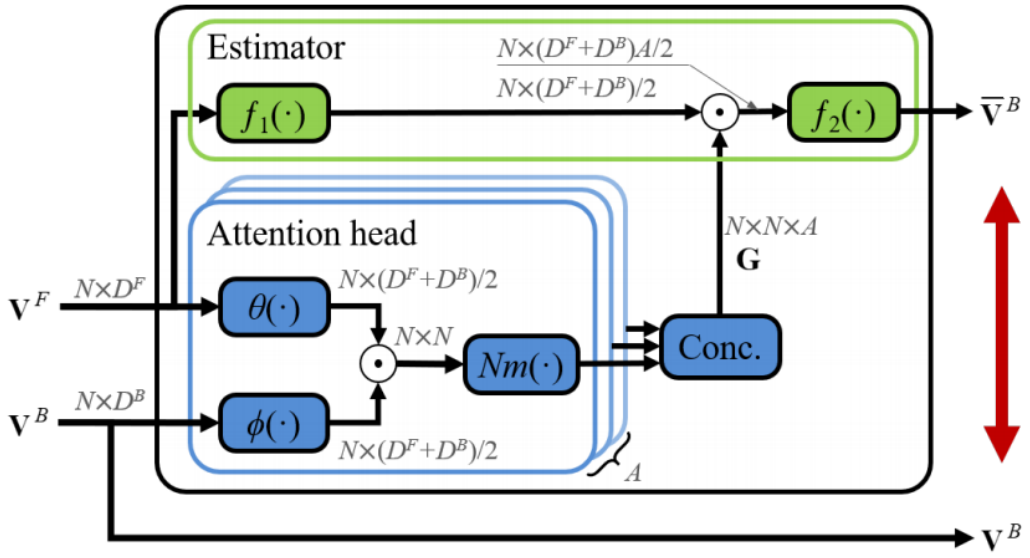
Our Research Results

- Attention heads을 통한 embedding procedure knowledge distillation [7]
  - Singular vectors를 이용하여 inter-data relation을 계산
    → Cosine similarity로 계산된 affinity matrix를 사용
  - Inter-data의 변화 과정을 CNN의 embedding procedure 정의
  - Neural network를 통해 teacher knowledge를 soften
    → attention heads를 사용



Our Research Results

- Attention heads을 통한 embedding procedure knowledge distillation
  - 이전 단계의 singular vector set으로 다음 단계의 singular vector set을 학습
    → Attention heads가 inter-data relation을 분석



Our Research Results

- Attention heads을 통한 embedding procedure knowledge distillation
  - 실험 결과

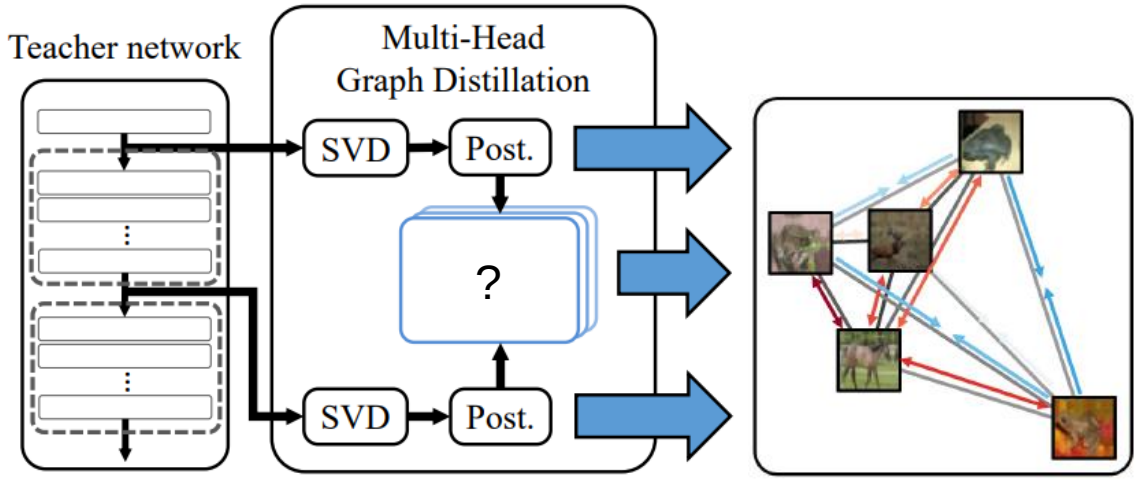
Dataset	Rate	Student	AT	FT	AB	RKD	MHGD
CIFAR100	Full	76.09	76.98	77.14	77.29	77.02	77.45
	0.50	69.77	71.13	72.41	72.28	69.57	73.32
	0.25	59.28	63.07	63.70	66.79	53.57	67.27
	0.10	40.65	47.66	48.29	57.38	23.27	54.58
TinyImageNet	Full	59.71	60.92	55.61	60.19	61.12	62.26
	0.50	52.53	54.50	55.81	54.41	54.09	56.56
	0.25	43.56	46.54	39.19	48.99	42.19	50.59
	0.10	28.44	32.38	34.08	42.18	20.90	38.28

Dataset	Rate	Student	AT	FT	AB	RKD	MHGD
CUB200-2011	Full	52.21	58.87	59.96	56.80	52.54	55.77
	0.50	30.58	39.51	42.94	39.77	29.72	34.02
	0.25	14.25	19.68	21.18	20.52	14.15	18.41
	0.10	5.87	8.05	8.04	7.03	6.60	5.97
MIT-scene	Full	51.00	56.32	60.07	59.52	53.50	47.90
	0.50	36.83	42.43	46.53	46.80	39.18	36.48
	0.25	21.59	28.54	31.96	33.13	25.39	25.51
	0.10	10.59	14.44	14.39	19.79	12.17	10.07

Our Research Results

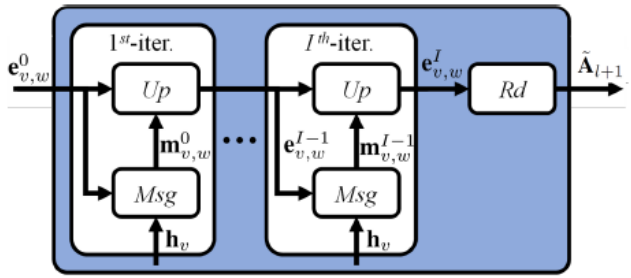
- 문제점 발생
  - 학습된 attention heads가 생성하는 graph가 너무 noisy함
    → 예상했던 것과 달리 사람이 이해할 수 없는 information을 추출..

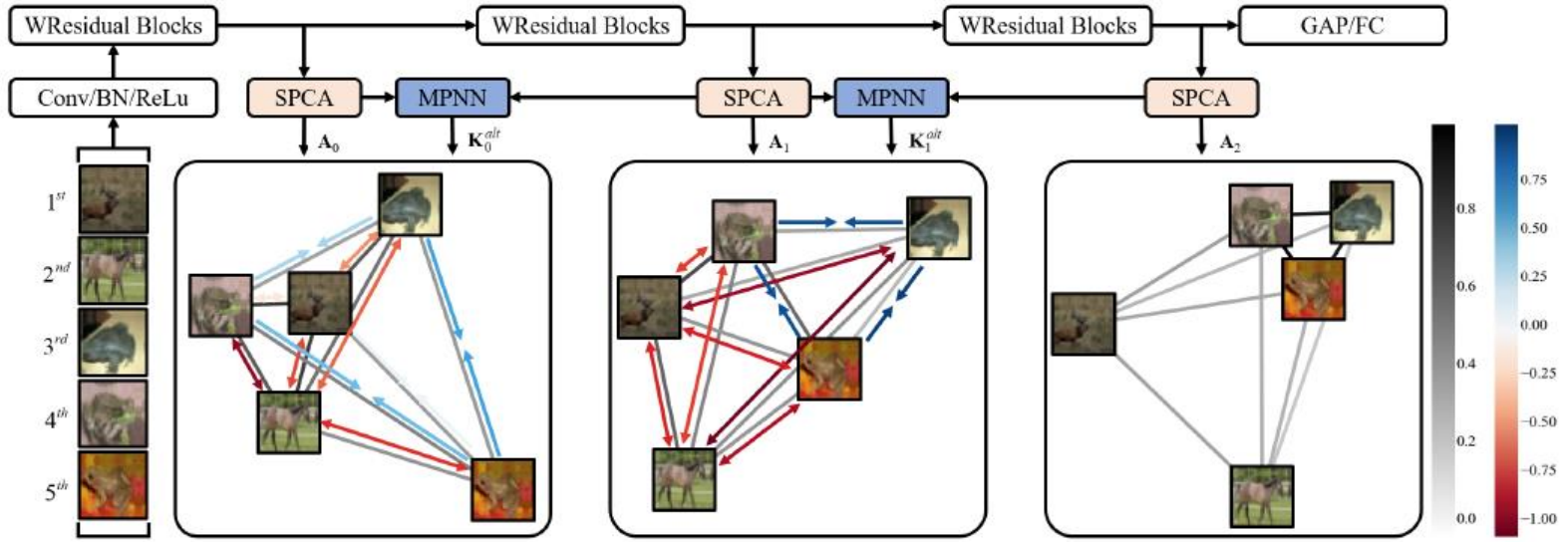
  - 사람이 이해할 수 있는 형태의 embedding procedure를 얻을 수 있을까?



Our Research Results

- MPNN을 통한 embedding procedure knowledge distillation [13]
  - MPNN을 통해 이전 단계의 graph로 다음 단계를 예측
    → MPNN의 message가 inter-data relation의 변화를 나타냄

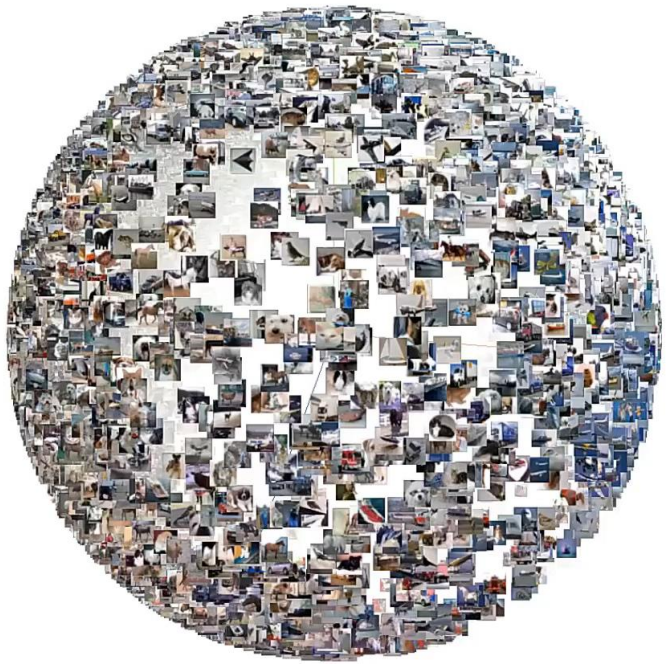




[13] Lee, Seunghyun, and Byung Cheol Song. "Interpretable Embedding Procedure Knowledge Transfer via Stacked Principal Component Analysis and Graph Neural Network." AAAI 2021.

Our Research Results

- MPNN을 통한 embedding procedure knowledge distillation
  - Visualization



Our Research Results

- IEPKT

Dataset	Rate	Student	AT	FT	AB	RKD	MHGD	CO	IEP	IEP+Black-box
CIFAR100	Full	76.09	76.98	77.14	77.29	77.02	77.45	78.21	78.12	**78.37**
	0.50	69.77	71.13	72.41	72.28	69.57	73.32	74.33	74.22	**74.53**
	0.25	59.28	63.07	63.70	66.79	53.57	67.27	67.90	68.57	**69.02**
	0.10	40.65	47.66	48.29	57.38	23.27	54.58	40.80	55.89	**59.04**
TinyImageNet	Full	59.71	60.92	55.61	60.19	61.12	62.26	63.56	63.29	**63.73**
	0.50	52.53	54.50	55.81	54.41	54.09	56.56	59.14	58.56	**59.27**
	0.25	43.56	46.54	39.19	48.99	42.19	50.59	52.56	53.20	**53.68**
	0.10	28.44	32.38	34.08	42.18	20.90	38.28	34.73	43.00	**45.01**

Table 1: Small network enhancement performance comparison of several KD methods for CIFAR100 and TinyImageNet datasets with various sample rates.

Dataset	Rate	Student	AT	FT	AB	RKD	MHGD	CO	IEP	IEP+Black-box
CUB200-2011	Full	52.21	58.87	59.96	56.80	52.54	55.77	60.83	60.13	**61.35**
	0.50	30.58	39.51	42.94	39.77	29.72	34.02	37.61	42.24	**43.06**
	0.25	14.25	19.68	21.18	20.52	14.15	18.41	14.29	22.00	**22.60**
	0.10	5.87	8.05	8.04	7.03	6.60	5.97	4.61	8.74	**9.69**
MIT-scene	Full	51.00	56.32	60.07	59.52	53.50	47.90	57.72	59.32	**60.94**
	0.50	36.83	42.43	46.53	46.80	39.18	36.48	35.16	45.83	**47.85**
	0.25	21.59	28.54	31.96	33.13	25.39	25.51	21.14	33.83	**34.28**
	0.10	10.59	14.44	14.39	19.79	12.17	10.07	6.07	18.44	**19.94**

Table 2: Transfer learning performance comparison of several KD methods for CUB-200-2011 and MIT-scene datasets with various sample rates.

Our Research Results

- Summary and conclusion
  - 어느 정도 사람이 납득할 수 있는 형태의 knowledge를 정의
  - SVD를 통한 feature maps의 압축
    → Feature map의 information을 유지하면서 이후 process의 cost를 효과적으로 감량
  - Embedding procedure는 유의미한 information을 가지고 있어 다양한 활용이 가능할 것으로 생각됨
  - 하지만, 기존 knowledge distillation 기법들에 비하면 너무 큰 cost를 요구..