Interpretable Embedding Procedure Knowledge Transfer via Stacked Principal Component Analysis and Graph Neural Network

Seunghyun Lee*, Byung Cheol Song

Department of Electrical and Computer Engineering, Inha University, Republic of Korea

Introduction

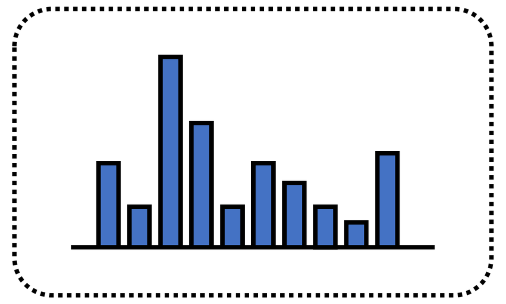
Knowledge Transfer
- Enhance shallow and simple network by transferring deep and complex network's knowledge.

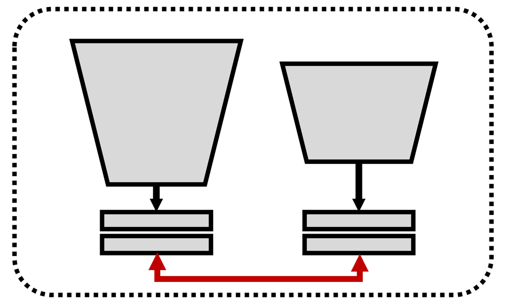
Problem Definition
- Conventional knowledge **doesn't coincide with CNN's goal** that is embedding datasets into low-dimensional space.
- It is mostly **hard to interpret its information**.

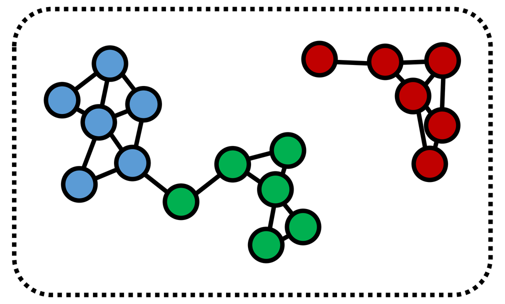
Contribution Points
- **Interpretable knowledge of embedding procedure**, which matches to human insight.
- **SOTA performance** by transferring CNN's complete knowledge.

Related works

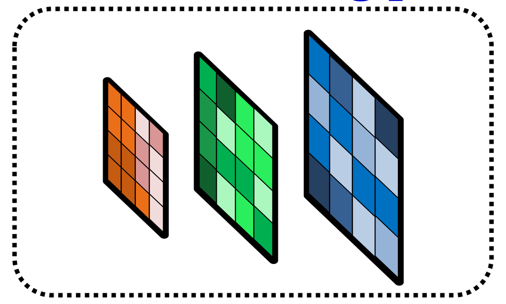
Embedded Feature Transfer Algorithm
- Extract information from better CNN's output.
- Some papers proposed the knowledge of inter-data relation, which is **limited to the embedded results**.

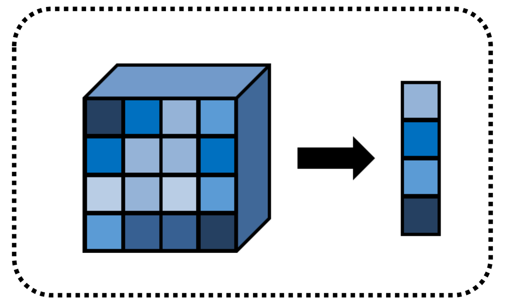

Softened output [1]

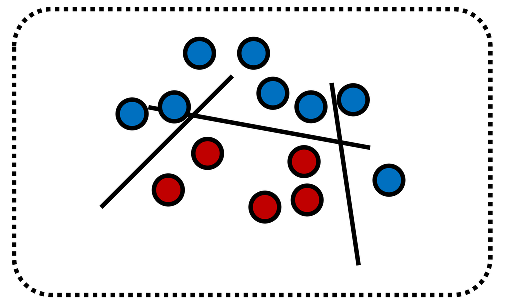

Factor transfer [2]


Inter-data relation [3]

Latent Feature Transfer Algorithm
- Extract information from multiple latent feature maps to increase the quantity of knowledge.
- Feature maps have quite complex information so that the derived knowledge is **not interpretable and has a far distance to the embedding procedure**.


Attention transfer [4]


Singular vectors [5]


Activation boundary [6]

References

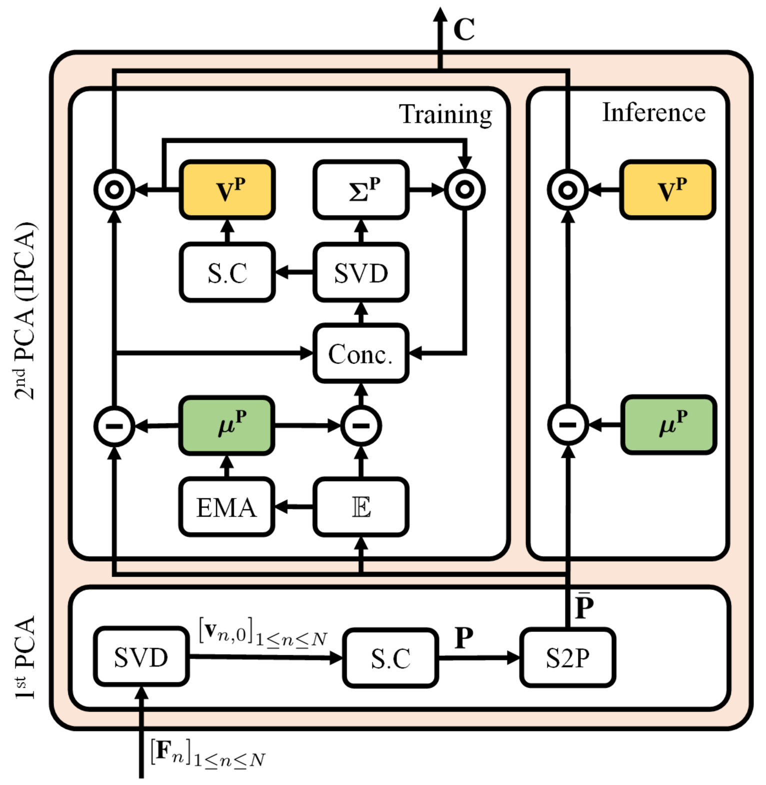
[1] Distilling the knowledge in a neural network. NIPS 2014 Deep Learning Workshop
[2] Paraphrasing Complex Network: Network Compression via Factor Transfer NeurIPS2018
[3] Relational Knowledge Distillation. CVPR2019
[4] Paying more attention to attention. ICLR2017
[5] Self-supervised knowledge distillation using singular value decomposition. ECCV 2018
[6] Knowledge transfer via distillation of activation boundaries formed by hidden neurons. AAAI2019
[7] Linguistically-Informed Self-Attention for Semantic Role Labeling. EMNLP 2018
[8] Graph-based Knowledge Distillation by Multi-head Attention Network. BMVC 2019

Method

Stacked Principal Component Analysis
- Compress feature maps to analyze the embedding procedure.
- 1st PCA: Compress a feature map into a principal component.
- 2nd PCA: Compress a principal component once more to increase compression rate and make it **available to be interpreted**.

※ Inter-data relation is captured by affinity matrices, which has information about the embedding procedure.

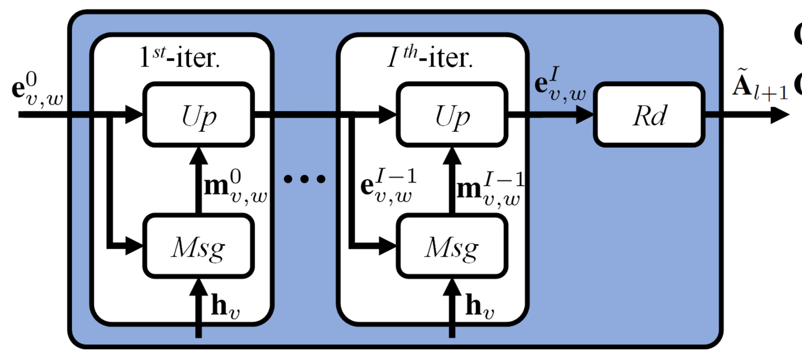
$$\mathbf{A} = \left[\frac{1}{\|\mathbf{c}_v\|_2 \|\mathbf{c}_w\|_2}\mathbf{c}_v^* \cdot \mathbf{c}_w\right]_{1\le v,w\le N}$$



Message Passing Neural Network
- Distill the embedding procedure knowledge through estimating the affinity matrices by message passing neural network.
- The message $\mathbf{m}^i_{v,w}$ updates a previous affinity matrix into the next affinity matrix $\tilde{\mathbf{A}}_{l+1}$.

$$\mathbf{e}^{i+1}_{v,w} = Up\left(\mathbf{e}^i_{v,w}, \boxed{\mathbf{m}^i_{v,w}}\right), \quad \boxed{\tilde{\mathbf{A}}_{l+1}} = \left[Rd\left(\mathbf{e}^I_{v,w}\right)\right]_{1\le v,w\le N}$$

※ The estimated affinity matrices contain the interim embedding knowledge $\mathbf{K}^{int}$ and the messages represent their alteration knowledge $\mathbf{K}^{alt}$.

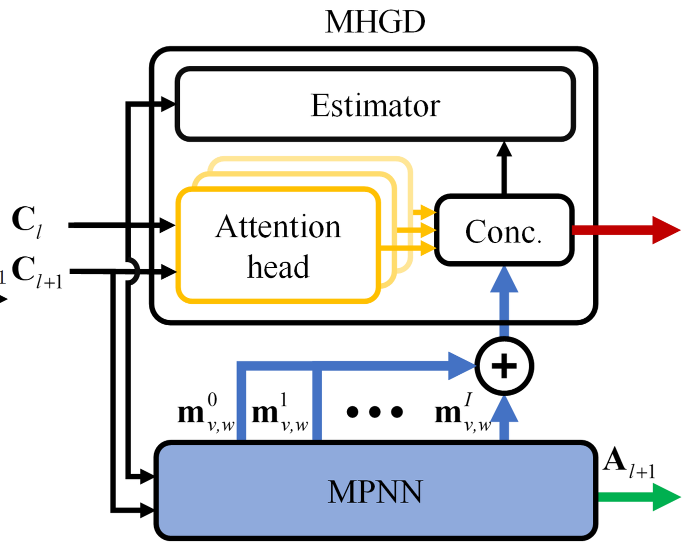


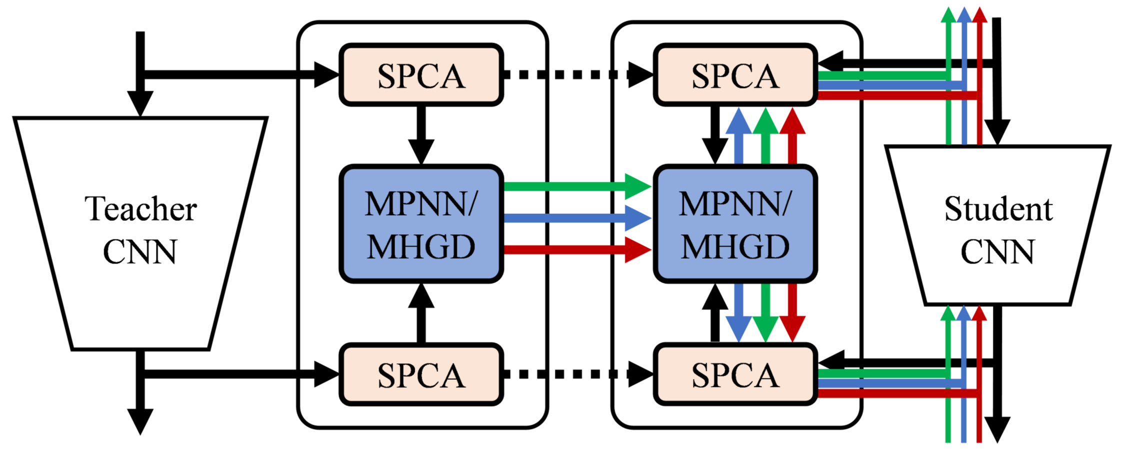
Transfer Knowledge with Gradient Clipping
- Two kinds of knowledge $\mathbf{K}^{int}$ and $\mathbf{K}^{alt}$ have different scales.
- In order to balance each knowledge's constraint, apply gradient clipping.

$$\frac{\partial\Theta}{\partial\mathcal{L}^{Total}} = \frac{\partial\Theta}{\partial\mathcal{L}^{Target}} + clip\left(\frac{\partial\Theta}{\partial\mathcal{L}^{int}}\right) + clip\left(\frac{\partial\Theta}{\partial\mathcal{L}^{alt}}\right) \qquad clip\left(z\right) = \max\left(1, \left\|\frac{\partial\Theta}{\partial\mathcal{L}^{Target}}\right\|_2 / \|z\|_2\right) z$$

Black-box Knowledge Distillation via Multi-head Graph Distillation
- CNN has a black-box which is not still interpretable.
- Adopt the concept of LISA [7] to distill the black-box knowledge, which accomplishes complete CNN's knowledge with IEP knowledge.

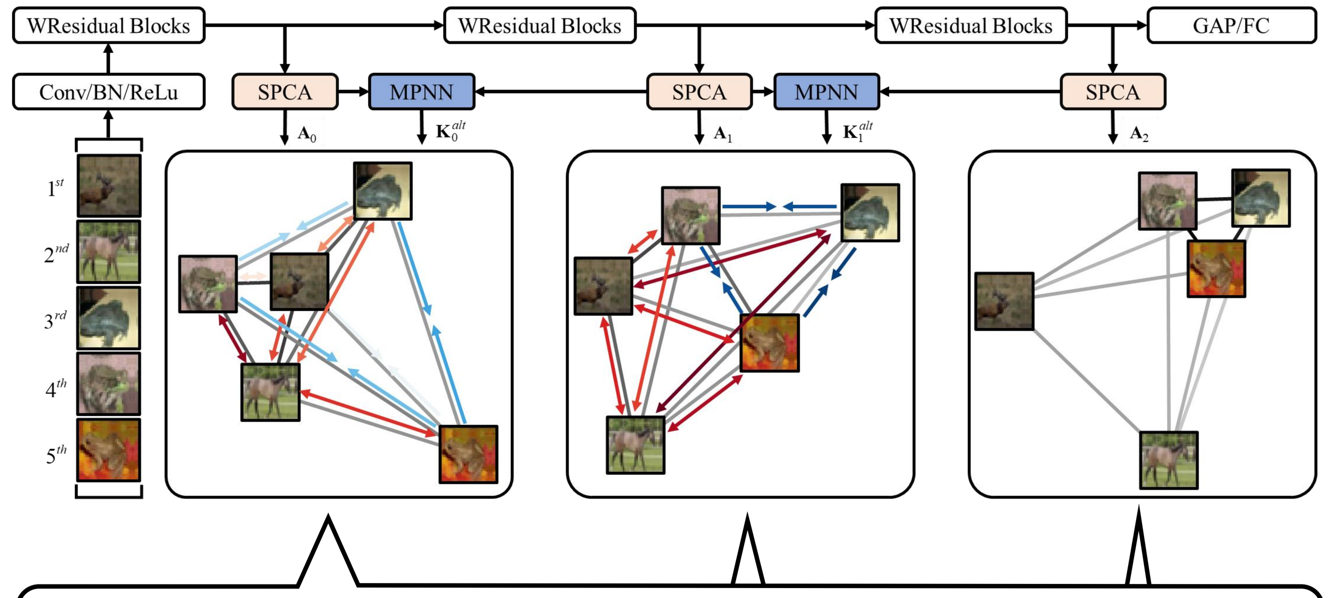
※ MHGD [8] is adopted as black-box knowledge distiller.





Experimental results

Visualization of IEP knowledge
- The proposed IEP knowledge successfully shows how CNN embeds a dataset into the label space.



- $\mathbf{A}_0$ shows that **similar colored images have a high-relationship**.
- $\mathbf{K}_0^{alt}$ shows that **images with the same class should get closer**.

- $\mathbf{A}_1$ and $\mathbf{K}_1^{alt}$ show a **similar aspect to the previous graph**.

- $\mathbf{A}_0$ shows that images are **clustered according to their classes**.

Small Network Enhancement
- The performance gaps increase as the sample rate decreases.

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**

Transfer Knowledge into different domain or architecture
- Our knowledge outperforms others because it represents not the feature map itself but an inter-data relation.

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**

Architecture	Student	AT	FT	AB	RKD	MHGD	CO	IEP+Black-box
WResNet16-2	56.61	59.42	57.28	62.53	54.27	59.29	60.21	**63.78**
WResNet16-1	51.88	53.01	50.95	55.01	48.46	50.72	52.67	**56.09**
MobileNet-V2	56.96	59.04	57.48	61.35	58.17	61.80	62.72	**64.82**
VGG	47.76	49.88	48.13	N/A	N/A	47.40	45.18	**55.82**

Ablation study
- Each knowledge gives sufficient performance gain.

Dataset	Student	$\mathbf{K}^{int}$	$\mathbf{K}^{alt}$	IEP	$\mathbf{K}^{BB}$
CIFAR100	59.28	61.77	68.14	68.57	67.75
TINY	43.56	45.90	52.62	53.26	51.92

- Incremental PCA better represents embedding spaces.

Sample rate	100%	50%	25%	10%
PCA-IPCA	78.12	74.22	68.57	55.89
PCA-PCA	77.92	73.66	66.58	51.88

- Too many iterations in MPNN gives over-constraints.

Iteration	1	2	3	4
CIFAR100	67.91	68.57	68.44	66.63
TinyImageNet	52.81	53.20	53.28	51.55

Conclusion
- The proposed knowledge gives not only SOTA performance gain but also tools for interpreting CNN's behavior.
- Need to more focus on what is CNN's authentic knowledge, not the performance.

* Code is available at https://github.com/sseung0703/IEPKT