Layer-Centric Factors of Variation Disentanglement for Task- and Model-Agnostic Generalization

Hee-Jun Jung, Jongmin Park, Minwoo Kang, Hoyong Kim, Kangil Kim

Gwangju Institute of Science and Technology (GIST)

ICML International Conference On Machine Learning

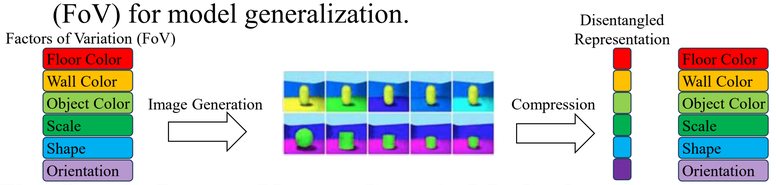
GIST Intelligence Representation &Reasoning Lab

Paper Code

Motivation

Disentanglement Learning

1. Identifying and Separating the underlying Factors of Variations (FoV) for model generalization.



The Gap in Latent-Vector-Centric Methods

1. Conventional disentanglement methods impose objectives only at the latent bottleneck.
2. Improved intrinsic metrics do not reliably translate into downstream generalization gains.
3. Downstream predictions are shaped by intermediate features throughout the network.

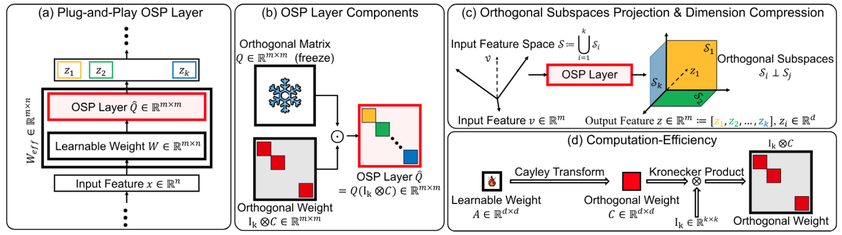
What We Target?

1. What?: Disentanglement should be induced at the layer level, not just the latent space.
2. How?: Given intermediate feature $v \in R^m$, separate it into K mutually orthogonal subspaces.
3. Why?: Forming an independent subspace structure that mirrors the common assumption of (approximately) independent FoV.

Contributions

1. Proposes a continuous layer-centric disentanglement method.
2. Provides a plug-and-play OSP layer for convolutional, linear, and Transformer-based architectures.
3. Demonstrates improvements in disentanglement and downstream generalization across CV, NLP, and fine-tuning tasks.

Method



1) Orthogonal Subspace Projection

1. Let $A' = QR, Q \in \mathbb{R}^{m \times m}, Q = [Q_1, Q_2, \ldots, Q_k], Q_i \in \mathbb{R}^{m \times d}$.
2. Then $P_i = Q_i Q_i^T \in \mathbb{R}^{m \times m}$ is the orthogonal projection onto the d-dimensional subspace $\mathcal{S}_i = col(Q_i)$.
3. Any vector $v \in \mathbb{R}^m$ projected onto subspace $\mathcal{S}_i$ and $\mathcal{S}_j$ is orthogonal $\boldsymbol{P_i v} \perp \boldsymbol{P_j v}$.

2) Dimension Compression

1. Generating the set of k projected vectors $\{p_i v\}_{i=1}^{k}$ **increase** memory usage.
2. Projection matrix rank (d) is **lower** than output dimension (m), where $rank(P_i) = d \ll m$.
3. Output feature $z = [z_1, z_2, \ldots z_k] \in \mathbb{R}^{kd} = \mathbb{R}^m$, where $z_i = Q_i^T v \in \mathbb{R}^d$.

3) Computational Efficiency

1. $\hat{Q} = [Q_1 C, Q_2 C, \ldots, Q_k C] = Q_{total}(I_k \otimes C)$.
2. $C = (I_d - A)(I_d + A)^{-1}, A^T = A^{-1}$.

4) Plug-and-Play

1. Linear and Convolutional Layers.

Results

Disentanglement Learning (DL)

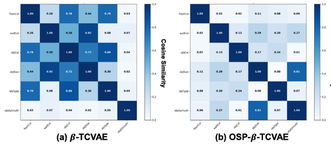
(1) OSP improves disentanglement without explicit disentanglement objectives.
(2) OSP-VAE further improves disentanglement performance.
(3) Factor directions become closer to the ideal identity-matrix structure.

Dataset	Metric	AE	OSP-AE
3D Shapes	β-VAE	67.20(±2.28)	**76.80**(±5.02)
	FVM	50.62(±4.15)	**60.73**(±7.00)
	MIG	5.46(±1.18)	**8.87**(±1.58)
	SAP	2.23(±0.23)	**2.93**(±0.40)
	DCI-Dis.	10.52(±2.09)	**14.66**(±2.69)
	DCI-Com.	11.34(±1.50)	**16.26**(±1.98)
MPI3D-toy	β-VAE	55.20(±5.22)	**61.33**(±4.62)
	FVM	33.48(±2.85)	**37.67**(±4.57)
	MIG	1.99(±0.38)	**2.65**(±1.61)
	SAP	1.23(±0.38)	**1.40**(±0.33)
	DCI-Dis.	7.70(±1.00)	**8.59**(±2.27)
	DCI-Com.	7.53(±1.03)	**8.18**(±2.11)

(1) AE vs. OSP-AE.

Dataset	Model	β-VAE	FVM	MIG
3D Shapes	β-VAE	85.80(±8.66)	75.86(±10.73)	35.41(±17.06)
	OSP-β-VAE	**90.60**(±4.72)	**78.28**(±10.06)	**43.33**(±13.95)
	β-TCVAE	88.80(±5.02)	74.70(±3.39)	34.27(±12.54)
	OSP-β-TCVAE	**90.80**(±5.22)	**84.88**(±4.64)	**41.32**(±14.06)
	CLG-VAE	79.60(±18.83)	68.80(±18.13)	34.41(±14.13)
	OSP-CLG-VAE	**85.80**(±8.30)	**78.68**(±10.59)	**34.79**(±12.42)
MPI3D	β-VAE	37.20(±16.39)	33.34(±10.90)	**6.91**(±9.78)
	OSP-β-VAE	**43.80**(±12.73)	**38.20**(±8.69)	5.83(±6.35)
	β-TCVAE	34.40(±15.19)	32.00(±12.35)	6.42(±11.09)
	OSP-β-TCVAE	**46.00**(±6.16)	**37.53**(±4.37)	**17.36**(±7.73)
	CLG-VAE	41.00(±5.03)	34.97(±1.04)	19.30(±1.78)
	OSP-CLG-VAE	**48.00**(±3.65)	**39.86**(±1.39)	**19.90**(±1.09)

(2) DL performance on 3D Shapes and MPI3D.



(3) Orthogonality between factor directions.

Computer Vision (CV)

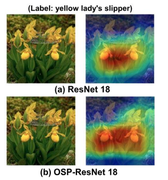
(4-5) OSP improves CV classification in both scratch training and fine-tuning.
(6) OSP-ResNet18 attends more to class-relevant object.

Models	Params	CIFAR-100	ImageNet
ResNet-18	11.69M	72.97$_{\pm 1.38}$	64.44$_{\pm 0.37}$
OSP-ResNet-18	11.70M	**74.00**$_{\pm 0.35}$	**65.71**$_{\pm 0.23}$
ResNet-50	25.56M	73.28$_{\pm 0.61}$	69.86$_{\pm 0.12}$
OSP-ResNet-50	25.58M	**74.64**$_{\pm 0.68}$	**70.79**$_{\pm 0.14}$
ViT-Base/16	87.40M	-	78.22$_{\pm 0.45}$
OSP-ViT-Base/16	87.58M	-	**80.33**$_{\pm 0.13}$

(4) Training from scratch

Models	Fine-Tuning	# of Param. ↓	ImageNet ↑
ResNet 18	Full fine-tuning	11.69M	66.69(±0.06)
	OPT (Liu et al., 2020)	46.50M	OOM
	S-OPT (Liu et al., 2020)	3.39M	67.74
	S-OPT	1.80M	50.07(±3.08)
	OSP-fine-tuning	0.50M	**67.78**(±0.08)
ResNet 50	Full fine-tuning	25.56M	76.59(±0.08)
	OSP-fine-tuning	1.00M	**77.67**(±0.08)

(5) Fine- tuning



(6) Visualization

Natural Language Processing

OSP improves word analogy and text classification performance.

	SAT			U2			U4			Google			BATS			Avg.		
	ppl	mppl	pmi	ppl	mppl	pmi	ppl	mppl	pmi	ppl	mppl	pmi	ppl	mppl	pmi	ppl	mppl	pmi
BERT	29.23	**28.76**	28.55	35.71	34.52	**37.50**	**37.50**	35.03	**38.13**	34.31	33.27	33.45	34.73	**35.04**	**36.46**	34.30	33.32	**34.82**
OSP-BERT	**30.43**	28.33	**29.97**	**39.29**	**39.29**	36.90	34.58	**36.67**	36.25	**34.52**	**34.89**	**34.52**	**43.76**	34.44	34.43	**34.71**	**34.73**	34.42
RoBERTa	**30.51**	29.03	29.86	36.76	34.33	35.71	34.43	35.21	35.83	33.06	**33.41**	32.83	33.25	33.25	**33.87**	33.60	33.04	33.86
OSP-RoBERTa	29.95	**30.19**	**31.99**	**39.48**	**37.90**	**37.90**	**38.75**	**38.33**	**38.96**	**33.21**	32.85	**33.77**	**33.97**	**34.16**	33.66	**35.07**	**34.69**	**35.26**

	Param.	WNLI	RTE	STS-B	MRPC	CoLA	SST-2	QNLI	QQP	MNLI
BERT	109.48M	39.44	60.29	86.29/86.08	81.37/87.21	**53.91**	92.32	90.66	90.51/87.31	83.91/84.10
OSP-BERT	109.63M	**46.48**	**65.89**	**86.38/86.40**	**81.62/87.44**	53.39	**92.43**	**91.16**	**90.60/87.43**	**84.52/84.48**
BERT-Large	335.14M	**56.33**	63.54	87.38/87.49	81.62/87.09	57.82	92.78	**92.26**	91.31/88.40	86.34/86.44
OSP-BERT-Large	335.40M	**56.33**	**94.62**	**88.36/88.38**	**82.60/87.80**	**58.59**	**93.00**	92.22	**91.42/88.50**	**86.68/86.75**
RoBERTa-large	355.43M	**56.34**	**72.20**	91.09/91.17	87.99/91.51	**64.07**	95.99	94.33	91.75/89.09	90.74/90.48
OSP-RoBERTa-large	355.76M	**56.34**	**72.20**	**91.25/91.34**	**88.73/91.90**	63.99	**96.33**	**94.42**	**91.91/89.30**	**90.98/90.76**