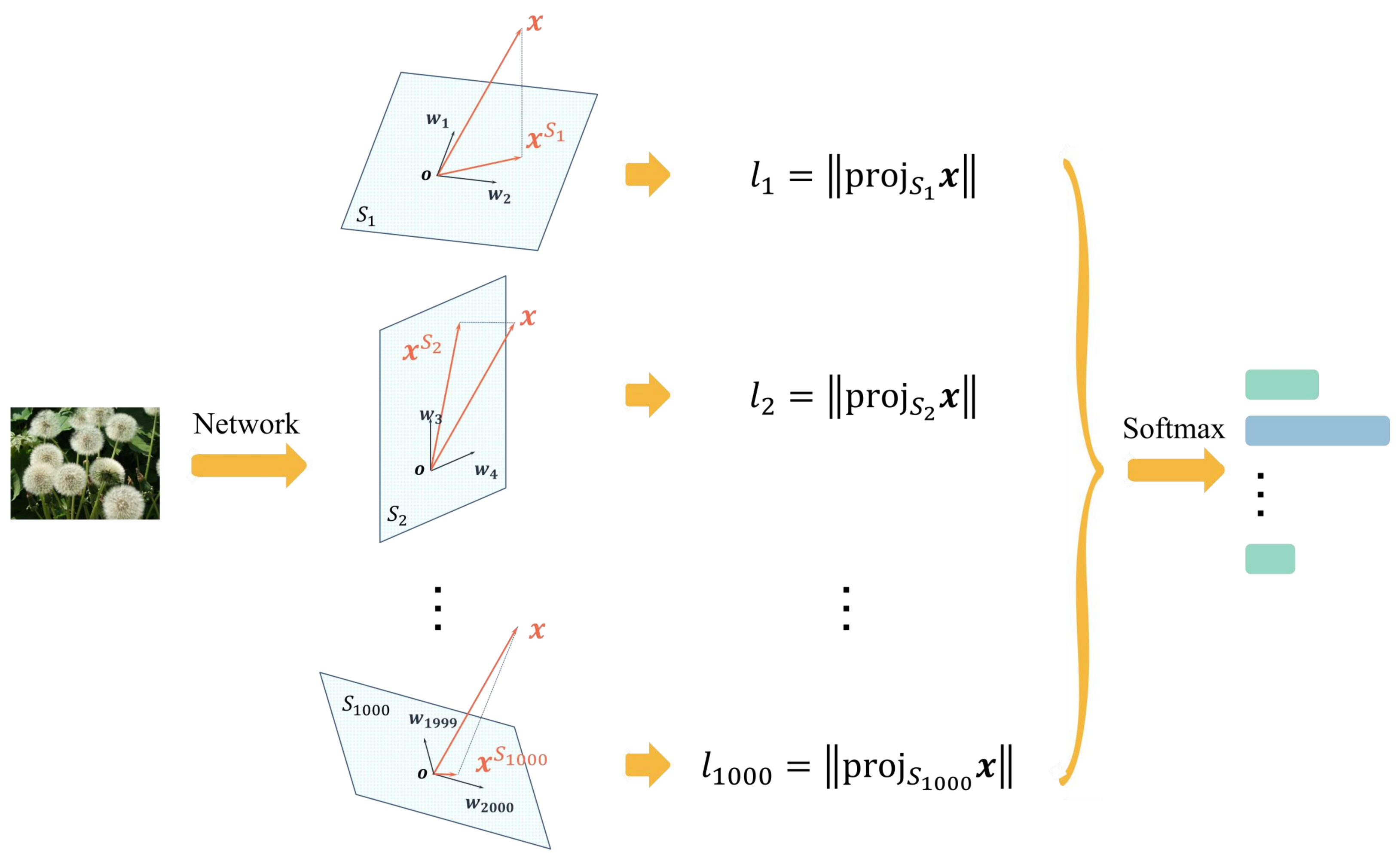
Get the Best of Both Worlds: Improving Accuracy and Transferability by Grassmann Class Representation

Haoqi Wang* Zhizhong Li* Wayne Zhang

EPFL sensetime ICCV23 PARIS

TL;DR Represent a class by a subspace in classification

Grassmann Class Representation



Grassmann Class Representation is to represent classes as linear subspaces in classification. The definition of **logit** is

$$l_i = \|\mathrm{proj}_{S_i}\boldsymbol{x}\|,$$

where S_i is a linear subspace representing the i-th class. Numerically, S_i is written as a matrix consisting of its **orthonormal bases**.

Motivation

- Hypothesis of **neural collapse**: features will reach minimal intra-class variability and maximal inter-class separability.
- Observation in recent literature: the **collapse** of intra-class variability **hurts** performance of feature transfer.

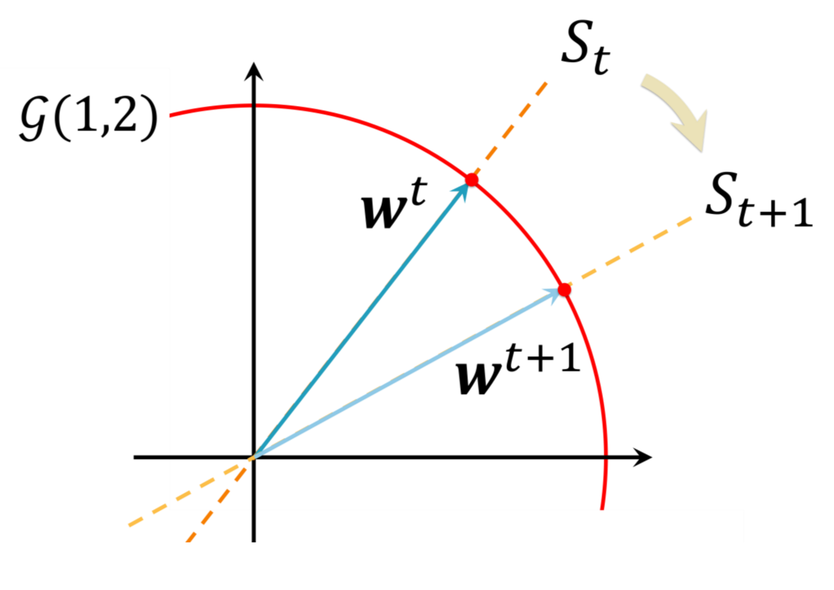
Question

How to **increase intra-class variability** while at the same time **maintain inter-class separability**?

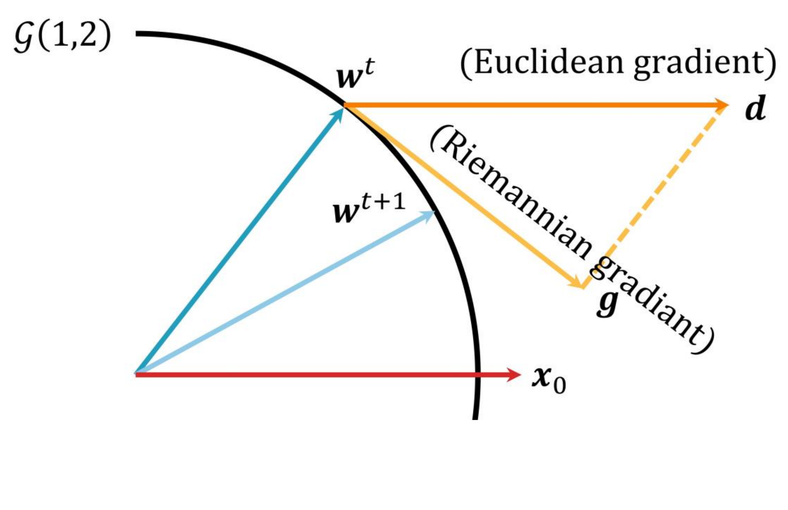
Answer

Model class as subspace. Allow features **vary within class**.

Riemannian Optimization



$\mathcal{G}(1,2)$: lines in 2d plane

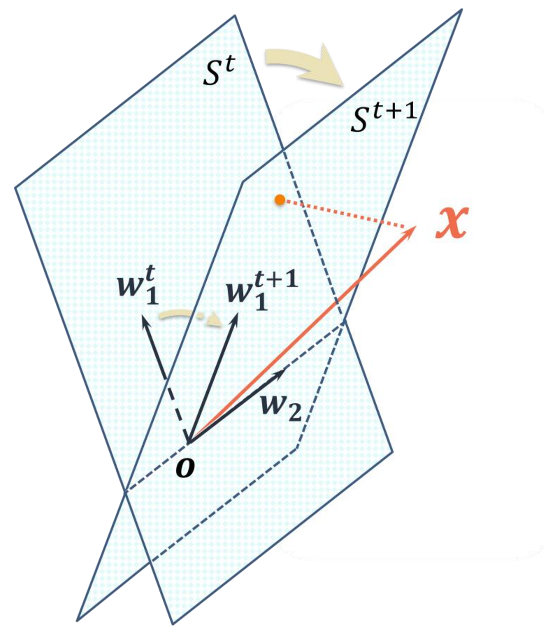


$$\max_{S\in\mathcal{G}(1,2)} \|\mathrm{proj}_S \boldsymbol{x}_0\|$$

The set of k-dim subspaces in n-dim Euclidean space form a **Grassmann manifold** $\mathcal{G}(k,n)$.

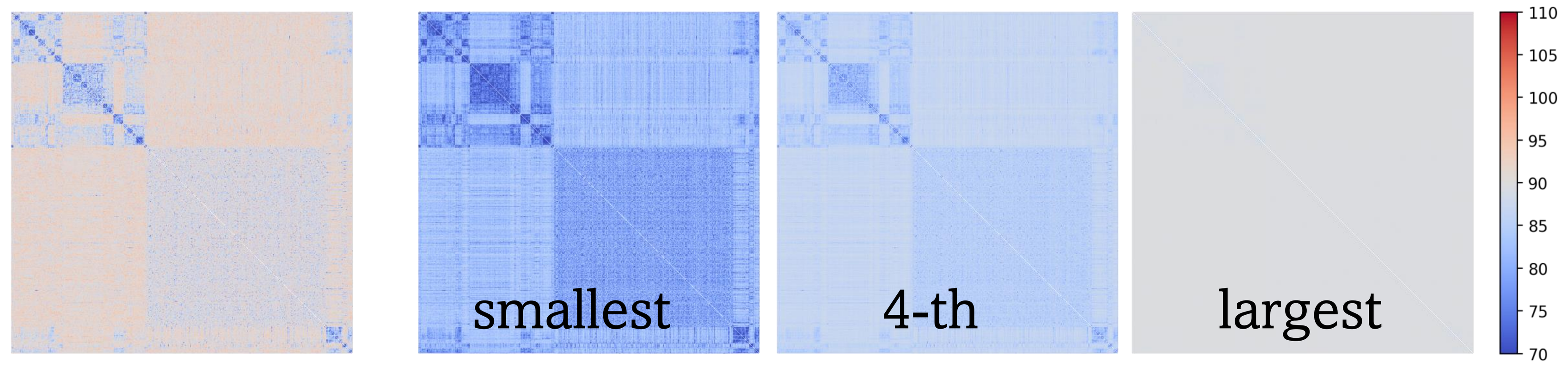
We use **Riemannian SGD** to optimize class subspaces.

1. Euclidean gradient to Riemannian gradient by projection
2. Update momentum
3. Move along geodesic toward gradient $\boldsymbol{G}$
   $\boldsymbol{S}(t) = (\boldsymbol{SV}\cos(t\boldsymbol{\Sigma}) + \boldsymbol{U}\sin(t\boldsymbol{\Sigma}))\boldsymbol{V}^T$,
   where $\boldsymbol{G} = \boldsymbol{U\Sigma V}^T$ is thin SVD.



Angles Between Classes

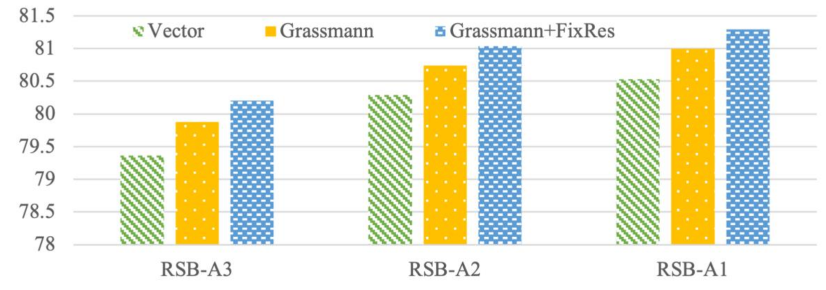
The pair-wise angles between 1000 classes of ImageNet-1K.



class as vector class as 8-dim subspace, principal angles

Classification Improvements

Setting	Top1	Top5	Class Representation
Softmax [8]	78.04	93.89	vector class representation
CosineSoftmax [19]	78.30	94.07	1-dim subspace
ArcFace [11]	76.66	92.98	1-dim subspace with margin
MultiFC	77.34	93.65	8 fc layers ensembled
SoftTriple [38]	75.55	92.62	8 centers weighted average
SubCenterArcFace [10]	77.10	93.51	8 centers with one activated
GCR (Ours)	**79.26**	**94.44**	8-dim subspace with RSGD



Different class representations trained on ImageNet-1K, ResNet50-D is used unless otherwise specified.

Setting					Vector Class Representation				Grassmann Class Representation ($k=8$)			
Architecture	n	BS	Epoch	Lr Policy	Loss	Optimizer	**Top1**	**Top5**	Loss	Optimizer	**Top1**	**Top5**
ResNet50 [16]	2048	256	100	Step	CE	SGD	76.58	93.05	CE	RSGD+SGD	**77.77**(↑1.19)	**93.67**(↑0.62)
ResNet50-D [17]	2048	256	100	Cosine	CE	SGD	78.04	93.89	CE	RSGD+SGD	**79.26**(↑1.22)	**94.44**(↑0.55)
ResNet101-D [17]	2048	256	100	Cosine	CE	SGD	79.32	94.62	CE	RSGD+SGD	**80.24**(↑0.92)	**94.95**(↑0.33)
ResNet152-D [17]	2048	256	100	Cosine	CE	SGD	80.00	95.02	CE	RSGD+SGD	**80.44**(↑0.44)	**95.21**(↑0.19)
ResNeXt50 [52]	2048	256	100	Cosine	CE	SGD	78.02	93.98	CE	RSGD+SGD	**79.00**(↑0.98)	**94.28**(↑0.30)
VGG13-BN [42]	4096	256	100	Step	CE	SGD	72.02	90.79	CE	RSGD+SGD	**73.40**(↑1.38)	**91.30**(↑0.51)
Swin-T [26]	768	1024	300	WarmCos	LS	AdamW	81.06	95.51	LS	RSGD+AdamW	**81.63**(↑0.57)	**95.77**(↑0.26)
Deit3-S [45]	384	2048	800	WarmCos	BCE	Lamb	81.53	95.21	CE	RSGD+Lamb	**82.18**(↑0.65)	**95.73**(↑0.52)

Different backbones and training schedules. Networks are trained on ImageNet-1K.

Feature Transfer Improvements

Table 3: Linear transfer using SVM for different losses. ResNet50-D is used as the backbone, and model weights are pre-trained on ImageNet-1K. *Variability* measures the intra-class variability, and R^2 measures class separation.

Setting		ImageNet		Analysis		Linear Transfer (SVM)						
Name	k	Top-1	Top-5	Variability	R^2	CIFAR10	CIFAR100	Food	Pets	Cars	Flowers	**Avg.**
Softmax [8]		78.04	93.89	60.12	0.495	90.79	67.76	72.13	92.49	51.55	93.17	77.98
CosineSoftmax [19]		78.30	94.07	56.87	0.528	89.34	65.32	64.79	91.68	43.92	87.28	73.72
LabelSmoothing [44]		78.07	94.10	54.79	0.577	89.14	63.22	66.02	91.72	43.58	91.01	74.12
Dropout [43]		77.92	93.80	55.40	0.565	89.27	64.33	66.74	91.38	43.99	88.59	74.05
Sigmoid [5]		78.04	93.81	60.20	0.491	91.09	69.26	71.71	91.98	51.75	92.86	78.11
GCR (Ours)	1	78.42	94.14	56.50	0.534	89.98	66.34	64.34	91.37	42.97	86.85	73.64
	4	78.68	94.32	61.48	0.459	90.56	67.45	67.58	91.37	50.24	90.08	76.21
	8	**79.26**	**94.44**	63.49	0.430	90.13	67.90	70.06	91.85	53.25	92.64	77.64
	16	79.21	94.37	65.79	0.395	91.09	69.58	71.28	91.99	55.93	93.80	78.95
	32	78.63	94.05	67.74	0.365	91.35	69.49	71.80	92.47	58.05	95.04	**79.70**

Takeaway

- GCR is an effective way to modeling classes as subspaces
- Riemannian SGD is effective to learn subspaces
- GCR enhances accuracy and transferability simultaneously