北京大學
PEKING UNIVERSITY

Longhui Yu
Peking University

—Do something really practical and valuable

Research Interest:

- ML for Healthcare; Trustworthy AI; Interactive ML

Brief

Research interests:

- ML for Healthcare
- Trustworthy AI
- Interactive ML

Short-term Goal:

- Publish paper on *Nature and its series*

Long-term Goal:

- Using ML&AI resolve important healthcare problem

Career Planning:

- Faculty in well-known university
- Researcher in Deepmind or Google Brain

I would fill the statistics and healthcare knowledge .

Generalizing and Decoupling Neural Collapse via Hyperspherical Uniformity Gap

北京大學 PEKING UNIVERSITY

Neural Collapse: Underlying geometric explanation for deep neural networks

(NC1) Within-class variability collapse[3]:

$$\boldsymbol{\Sigma}_B^{\dagger}\boldsymbol{\Sigma}_W \rightarrow \mathbf{0},$$

where † denotes the Moore-Penrose pseudoinverse.

(NC2) Convergence to Simplex ETF:

$$\frac{\langle \boldsymbol{\mu}_c - \boldsymbol{\mu}_G, \boldsymbol{\mu}_{c'} - \boldsymbol{\mu}_G \rangle}{\|\boldsymbol{\mu}_c - \boldsymbol{\mu}_G\|_2 \|\boldsymbol{\mu}_{c'} - \boldsymbol{\mu}_G\|_2} \rightarrow \begin{cases} 1, & c = c' \\ \frac{-1}{C-1}, & c \neq c' \end{cases}$$

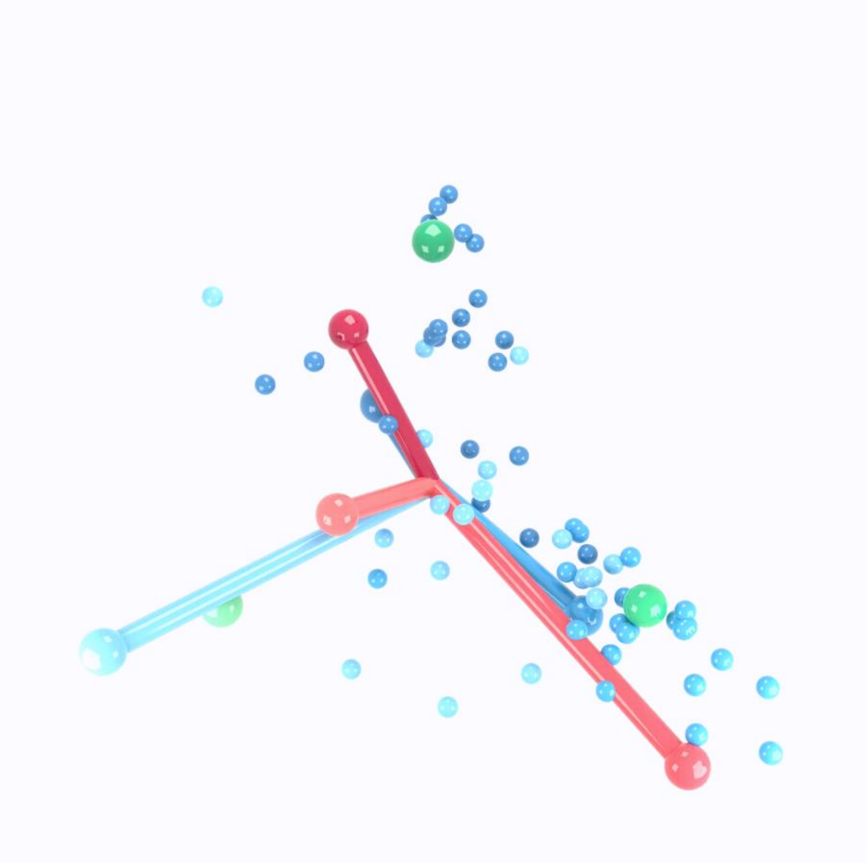
$$\|\boldsymbol{\mu}_c - \boldsymbol{\mu}_G\|_2 - \|\boldsymbol{\mu}_{c'} - \boldsymbol{\mu}_G\|_2 \rightarrow 0 \quad \forall c \neq c'$$

(NC3) Convergence to self-duality:

$$\frac{\boldsymbol{w}_c}{\|\boldsymbol{w}_c\|_2} - \frac{\boldsymbol{\mu}_c - \boldsymbol{\mu}_G}{\|\boldsymbol{\mu}_c - \boldsymbol{\mu}_G\|_2} \rightarrow 0$$

(NC4): Simplification to nearest class center:

$$\arg\max_{c'} \langle \boldsymbol{w}_{c'}, \boldsymbol{h} \rangle + b_{c'} \rightarrow \arg\min_{c'} \|\boldsymbol{h} - \boldsymbol{\mu}_{c'}\|_2$$



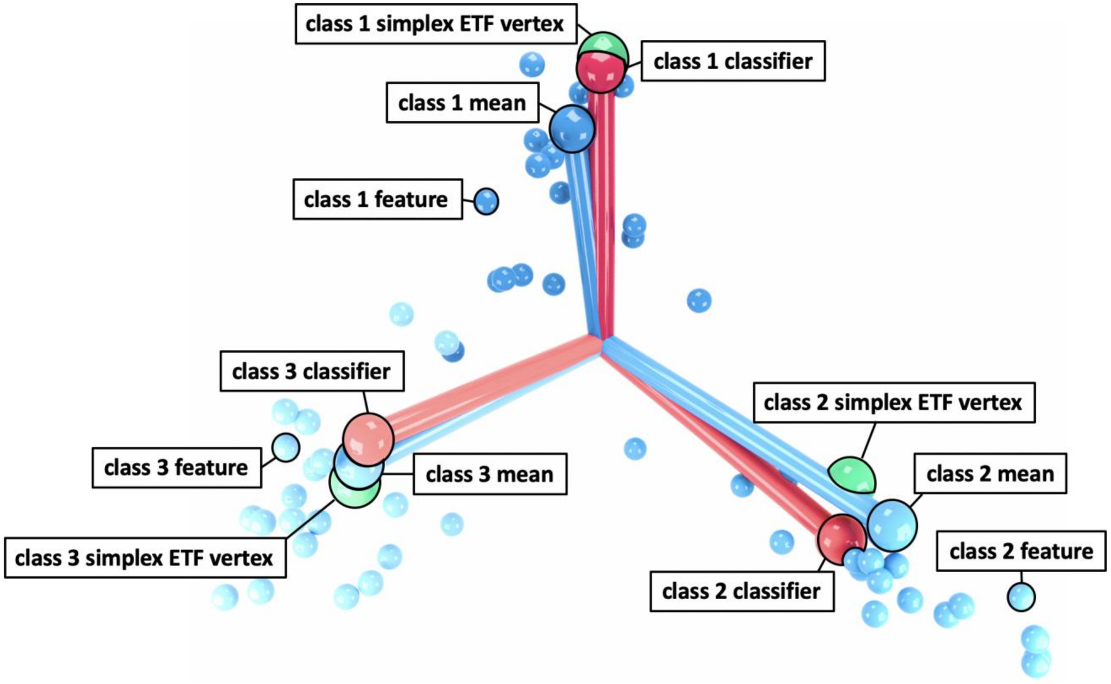
NC phenomenon suggests two general principles:

- **minimal intra-class compactness** of features
- **maximal inter-class separability** of classifiers / feature mean

Papyan V, Han X Y, Donoho D L. Prevalence of neural collapse during the terminal phase of deep learning training[J]. Proceedings of the National Academy of Sciences, 2020, 117(40): 24652-24663.

Neural Collapse

Underlying geometric explanation for deep neural networks



Papyan V, Han X Y, Donoho D L. Prevalence of neural collapse during the terminal phase of deep learning training[J]. Proceedings of the National Academy of Sciences, 2020, 117(40): 24652-24663.

Motivation: Decoupling

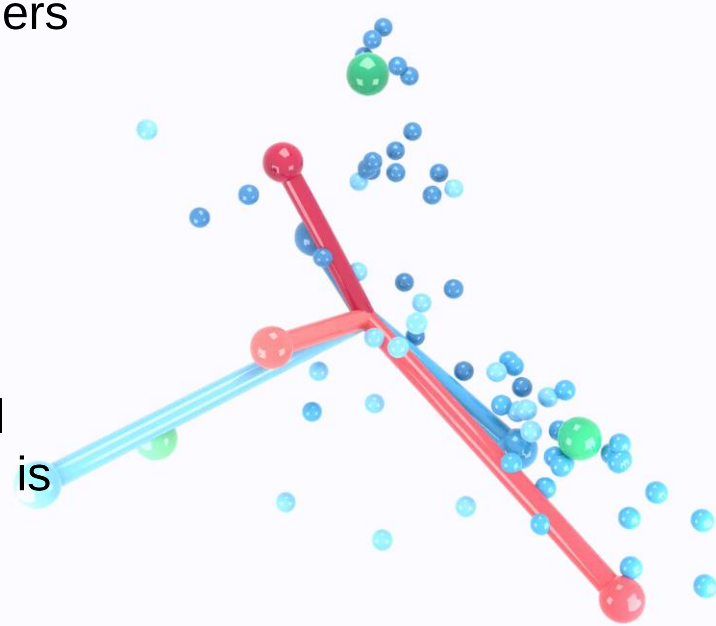
Popular loss like CE, MSE completely couple these two principles:

- **minimal intra-class compactness** of features
- **maximal inter-class separability** of classifiers / feature mean

For example:
You cannot optimize the two principles **independently** under CE loss.

To solve this, we propose HUG (hyperspherical uniformity gap) to substitute the CE loss, which is highly flexible.

This kinds of decoupling can make us analysis the influence of the optimization of each principle.



Papyan V, Han X Y, Donoho D L. Prevalence of neural collapse during the terminal phase of deep learning training[J]. Proceedings of the National Academy of Sciences, 2020, 117(40): 24652-24663.

Motivation: Optimize data directly

Projection FDA: $\max_{\boldsymbol{T}\in\mathbb{R}^{d\times r}} \operatorname{tr}\left(\left(\boldsymbol{T}^\top \boldsymbol{S}_w \boldsymbol{T}\right)^{-1} \boldsymbol{T}^\top \boldsymbol{S}_b \boldsymbol{T}\right)$ **Data FDA:** $\max_{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\in\mathbb{S}^{d-1}} \operatorname{tr}(\boldsymbol{S}_b) - \operatorname{tr}(\boldsymbol{S}_w)$

- between-class scatter matrix:

$$\boldsymbol{S}_w = \sum_{i=1}^{C} \sum_{j\in A_c} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^\top$$

- within-class scatter matrix:

$$\boldsymbol{S}_b = \sum_{i=1}^{C} n_i (\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})^\top$$

T can be seen as the deep neural parameters, which aims to linearize ***X*** in some nonlinear manifold space to the linear-separatable feature space.

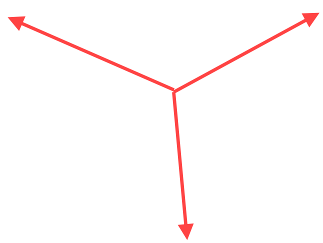
By utilizing the HUG to decouple the Neural Collapse, we can optimize the data directly. (note the data means feature before FC layer. This motivation is also used as analysis.)

Note: A concurrent work also analyze that: The solution of Neural Collapse can substitute the model solution for a analytical use.

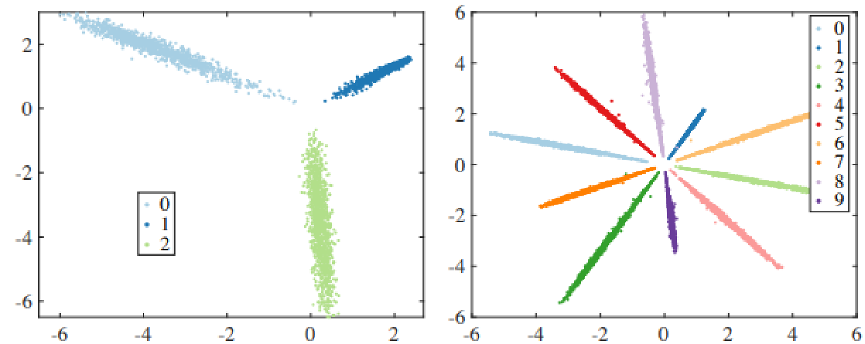
Motivation: Generalized Neural Collapse

Problems in Neural Collapse: In Neural Collapse, both features and classifiers converge to **ETF.** However, ETF exists when ***d*** $\geq$ ***C – 1.***
d: feature dimension. C: class number.
What will happen if d $<$ ***C – 1?***

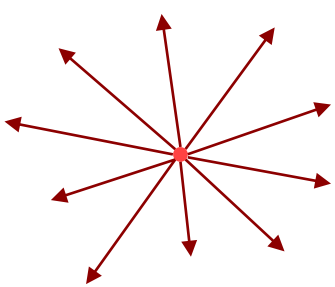
Interestingly, We find in this case, it lead to **hyperspherical uniformity**.



ETF & Uniformity



(a) 2D feature with 3 classes (b) 2D feature with 10 classes



Uniformity

Let the Uniformity to represent the Generalized Neural Collapse in all the cases.

However, can we represent both the two principles by Uniformity ???

We can !!!

HUG: General Framework and variants

$$\max_{\{\hat{\boldsymbol{x}}_i\}_{i=1}^n} \mathcal{L}_{\text{HUG}} := \alpha \cdot \underbrace{\mathcal{HU}\big(\{\hat{\boldsymbol{\mu}}_c\}_{c=1}^C\big)}_{T_b\text{: Inter-class Hyperspherical Uniformity}} - \beta \cdot \sum_{c=1}^{C} \underbrace{\mathcal{HU}\big(\{\hat{\boldsymbol{x}}_i\}_{i \in A_c}\big)}_{T_w\text{: Intra-class Hyperspherical Uniformity}}$$

In general, HUG aims to independently optimize the two principles:

T_b : Inter-class Hyperspherical Uniformity

T_w : Intra-class Hyperspherical Uniformity

In implementation, HUG should a proxy to do classifier-feature matching:

$$\max_{\{\hat{\boldsymbol{x}}_i\}_{i=1}^n, \{\hat{\boldsymbol{w}}_c\}_{c=1}^C} \mathcal{L}_{\text{P-HUG}} := \alpha \cdot \underbrace{\mathcal{HU}\big(\{\hat{\boldsymbol{w}}_c\}_{c=1}^C\big)}_{\text{Inter-class Hyperspherical Uniformity}} - \beta \cdot \sum_{c=1}^{C} \underbrace{\mathcal{HU}\big(\{\hat{\boldsymbol{x}}_i\}_{i \in A_c}, \hat{\boldsymbol{w}}_c\big)}_{\text{Intra-class Hyperspherical Uniformity}}$$

Variants:

- Minimum hyperspherical ener (MHE-HUG)
- Maximum hyperspherical separation (MHS-HUG)
- Maximum gram determinan (MGD-HUG)

Performance:

Method	CIFAR-10	CIFAR-100
CE Loss	5.45	24.90
MHE-HUG	**5.03**	**23.50**
MHS-HUG	5.09	24.38
MGD-HUG	5.38	24.59

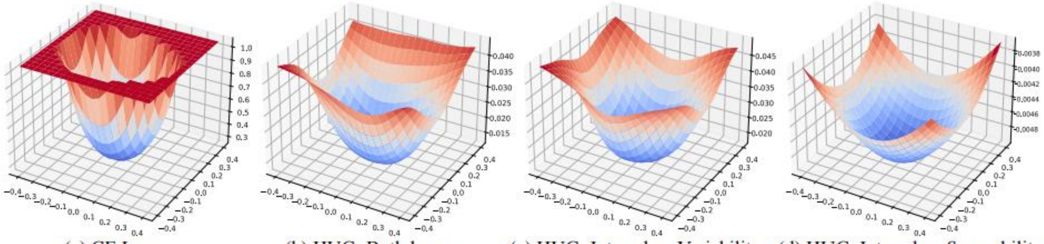
Table 1: Testing error (%) of HUG variants on CIFAR-10 and CIFAR-100.

Method	ResNet-18	VGG-16	DenseNet-121
CE Loss	5.45 / 24.90	5.28 / 22.99	5.04 / 21.47
HUG	**5.03 / 23.50**	**5.19 / 22.77**	**4.85 / 21.30**

Table 3: Testing error (%) with different architectures.

Loss landscape and convergence:

- **Interesting！ HUG can produce more flatten loss minima, which is better.**



(a) CE Loss (b) HUG: Both losses (c) HUG: Intra-class Variability (d) HUG: Inter-class Separability

Figure 4: Loss landscape visualization. (b,c,d) show $\mathcal{L}'_{\text{MHE-HUG}}$, T_b and T_w, respectively.

PEKING UNIVERSITY

Generalization:

Long-tailed recognition:

	CIFAR-100				CIFAR-10			
IR	0.2	0.1	0.02	0.01	0.2	0.1	0.02	0.01
CE	66.74	62.31	48.79	43.82	90.29	87.85	79.17	74.11
HUG	**67.83**	**63.33**	**50.48**	**45.63**	**90.41**	**88.20**	**79.88**	**75.14**

Table 4: Testing accuracy (%) of long-tailed recognition.

Continual Learning:

	CIFAR-100			CIFAR-10		
Memory size	200	500	2000	200	500	2000
ER + CE	22.14	31.02	43.54	49.07	61.58	76.89
ER + HUG	**23.52**	**31.92**	**43.92**	**53.74**	**62.67**	**77.21**

Table 5: Final testing accuracy (%) of continual learning.

Adversarial robustness:

Method	Clean	l_∞=2/255	l_∞=4/255	l_∞=8/255
CE Loss	5.45 / 24.90	7.94 / 2.12	0.61 / 0	0 / 0
HUG	**5.03 / 23.50**	**15.24 / 5.26**	**3.45 / 1.24**	**1.76 / 0.44**

Table 6: Testing accuracy (%) under adversarial attacks.

北京大學
PEKING UNIVERSITY

My Continual Learning Research

Longhui Yu

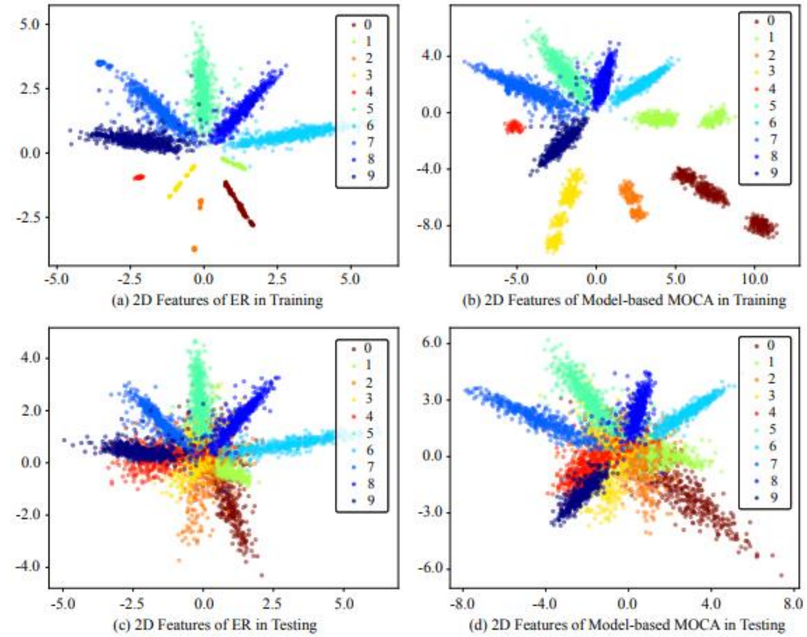
- Exploring the complex category relation in Continual Learning (ICME 2022 Oral)
- Taking full advantage of memory for Continual Learning (ICLR 2022)
- Designing a unified framework for Continual Learning (TMLR submit)

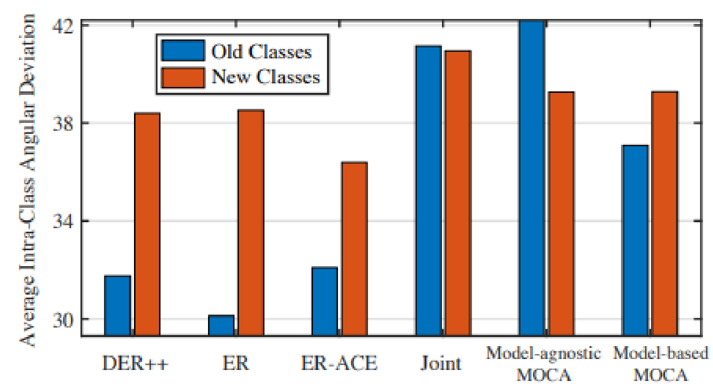
Continual Learning by Modeling Intra-Class Variation (Submission to TMLR)

北京大學 PEKING UNIVERSITY

Motivation:

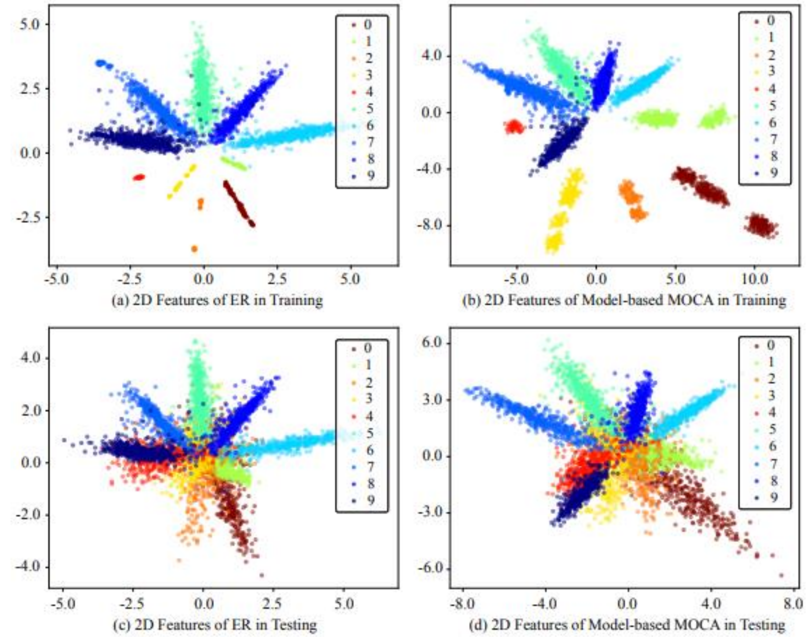
- **Due to the sample diversity, representational variation** is significantly different between old class and new class.
- ***The variation of old class representation is too small***. The old class feature collapse in a line. (a dot in a hypersphere)
- By adding ***MOCA***, the representation of old class can be diverse.
- By adding ***MOCA***, the variation gap between old class and new class can be reduced.



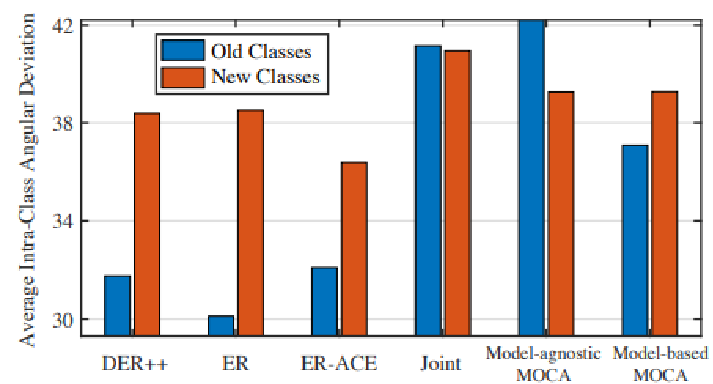


Intuition

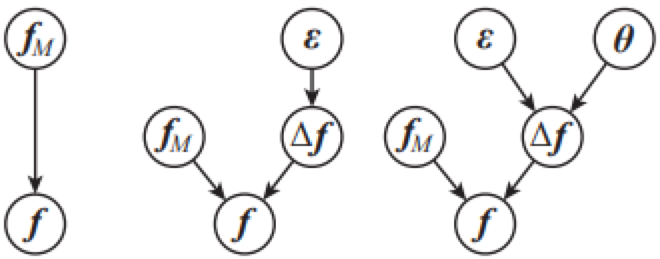
- The sample number of old class is significantly less than new class, For example, maybe 20 : 500.
- This inevitably cause the representation diversity gap.



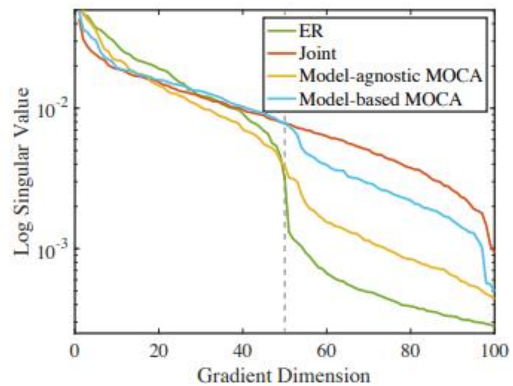
(a) 2D Features of ER in Training
(b) 2D Features of Model-based MOCA in Training
(c) 2D Features of ER in Testing
(d) 2D Features of Model-based MOCA in Testing



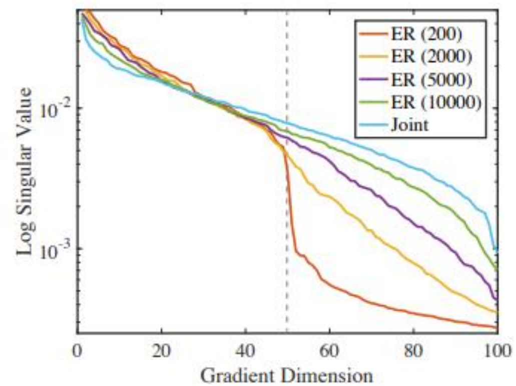
Motivation



(a) Original (b) Model-agnostic MOCA (b) Model-based MOCA



(a) Comparison of Different Methods



(b) Comparison of Different Memory Size

- MOCA as a gradient compensation method:
- For any continual learning method, if we can recover the training gradient under the joint training, we can recover the performance under the joint training.
- Experiments show Our MOCA achieve it successfully.

MOCA Framework

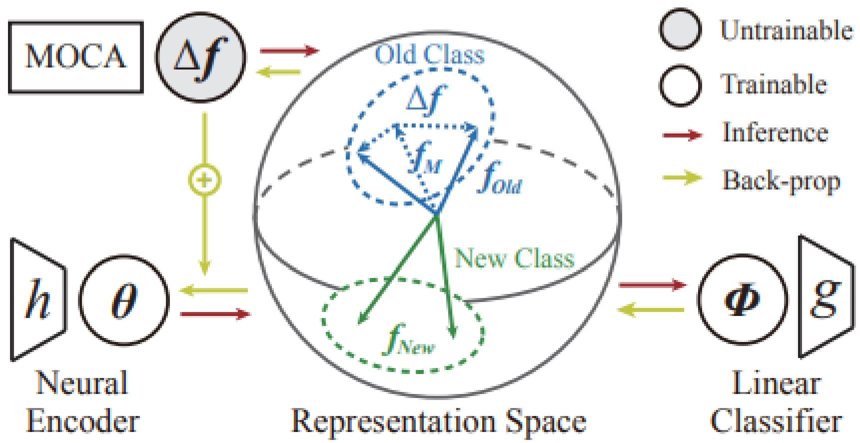


Figure 6: Inference and back-prop in MOCA.

- MOCA serves as a representation augmentation method:

$$\underbrace{\boldsymbol{f}}_{\text{Augmented Feature}} = \underbrace{h_{\boldsymbol{\theta}}(\boldsymbol{x})}_{\text{Prototype Feature}} + \underbrace{\left(\left(\|h_{\boldsymbol{\theta}}(\boldsymbol{x})\| - \|h_{\boldsymbol{\theta}}(\boldsymbol{x}) + \tilde{\Delta}\boldsymbol{f}\| \right) h_{\boldsymbol{\theta}}(\boldsymbol{x}) + \|h_{\boldsymbol{\theta}}(\boldsymbol{x})\| \, \tilde{\Delta}\boldsymbol{f} \right) \|h_{\boldsymbol{\theta}}(\boldsymbol{x}) + \tilde{\Delta}\boldsymbol{f}\|^{-1}}_{\text{Hyperspherical Augmentation } \Delta\boldsymbol{f}},$$

MOCA Framework

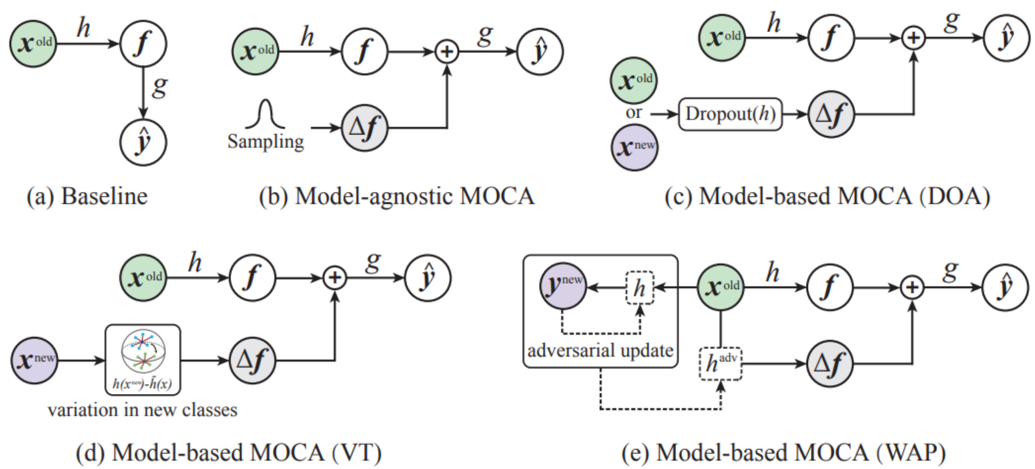
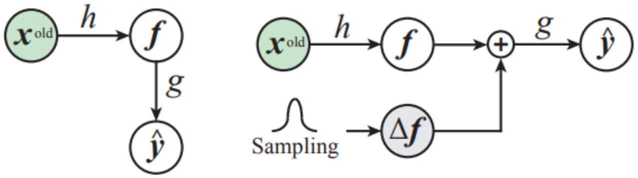


Figure 7: Illustration of different MOCA variants.

- The perturbation Δf can produced by two ways:
- Produced by a Probability Distribution. For example, Gaussian distribution, vMF distribution. This kind of method calls ***Mode-agnostic MOCA***.

- As the model feature space is in a high-dimensional manifold space. Considering the model knowledge to produce Δf is better, named as ***Mode-based MOCA***.

Model-agnostic MOCA



(a) Baseline (b) Model-agnostic MOCA

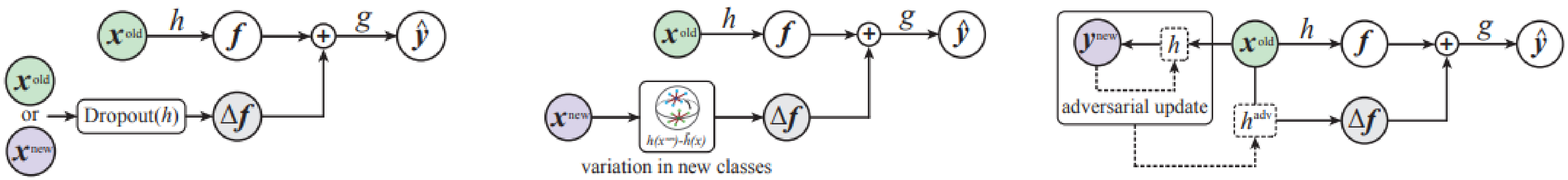
- **Isotropic Gaussian distribution:**

$$\boldsymbol{f} = \left\|h_{\boldsymbol{\theta}}(\boldsymbol{x}^{\mathrm{old}})\right\| \cdot \mathcal{P}_{\mathbb{S}}\big(\mathcal{P}_{\mathbb{S}}(h_{\boldsymbol{\theta}}(\boldsymbol{x}^{\mathrm{old}})) + \lambda \cdot \boldsymbol{\epsilon}\big),$$

- **von Mises–Fisher distribution:**

$$p(\boldsymbol{\epsilon}|\boldsymbol{\mu}, \kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)} \exp(\kappa \boldsymbol{\mu}^\top \boldsymbol{\epsilon}), \quad \boldsymbol{\mu} = \mathcal{P}_{\mathbb{S}}\big(h_{\boldsymbol{\theta}}(\boldsymbol{x}^{\mathrm{old}})\big),$$

Model-based MOCA



(c) Model-based MOCA (DOA) (d) Model-based MOCA (VT) (e) Model-based MOCA (WAP)

- **DOA:**

$$\boldsymbol{f}=\left\|h_{\boldsymbol{\theta}}(\boldsymbol{x}^{\text{old}})\right\|\cdot\mathcal{P}_{\mathbb{S}}\left(\mathcal{P}_{\mathbb{S}}\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}^{\text{old}})\right)+\lambda\cdot\mathcal{P}_{\mathbb{S}}\left(h_{\text{Dropout}(\boldsymbol{\theta})}(\boldsymbol{x})\right)\right),$$

- **VT:**

$$\boldsymbol{f}=\left\|h_{\boldsymbol{\theta}}(\boldsymbol{x}^{\text{old}})\right\|\cdot\mathcal{P}_{\mathbb{S}}\left(\mathcal{P}_{\mathbb{S}}\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}^{\text{old}})\right)+\lambda\cdot\mathcal{P}_{\mathbb{S}}\left(\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}^{\text{new}})-h_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}^{\text{new}})\right)\right)\right),$$

- **WAP:**

$$\boldsymbol{f}=\left\|h_{\boldsymbol{\theta}}(\boldsymbol{x}^{\text{old}})\right\|\cdot\mathcal{P}_{\mathbb{S}}\left(\mathcal{P}_{\mathbb{S}}\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}^{\text{old}})\right)+\lambda\cdot\mathcal{P}_{\mathbb{S}}\left(h_{\boldsymbol{\theta}+\Delta\boldsymbol{\theta}}(\boldsymbol{x})\right)\right),\ \text{s.t.}\ \Delta\boldsymbol{\theta}=\arg\min_{\|\Delta\boldsymbol{\theta}\|\le\epsilon}\mathcal{L}_{\text{ce}}\left(g_{\boldsymbol{\phi}}\left(h_{\boldsymbol{\theta}+\Delta\boldsymbol{\theta}}(\boldsymbol{x}^{\text{old}})\right),y^{\text{new}}\right),$$

- ***Model-based MOCA*** all consider the model θ to produce perturbation adding on the spherical feature

Experiments

Exhaustive experiments:

- ***Model-based MOCA*** performs better than ***Model-agnostic MOCA.***
- ***WAP (introducing adversarial attack to find the most useful perturbation)*** performs best among all of our approach.

Setting	Baseline	Gaussian	vMF	DOA-old	DOA-new	VT	WAP
Offline	31.08	37.29	**38.76**	33.67	38.75	39.78	**41.02**
Online	31.90	**32.78**	31.25	30.20	29.48	32.55	**33.72**
Proxy	31.26	**42.54**	42.24	-	45.72	**46.77**	-

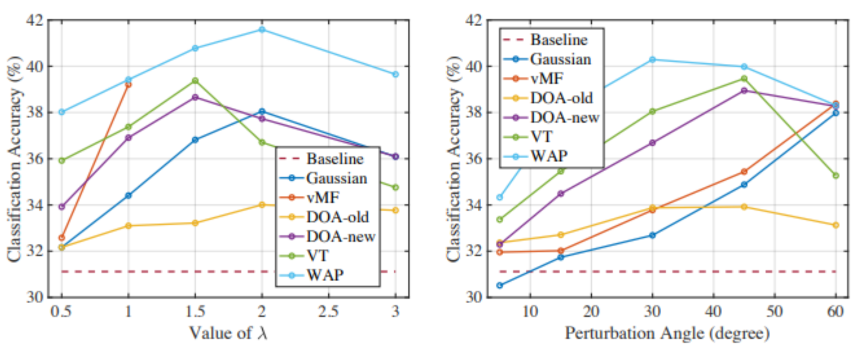


Figure 8: Left: the hyperparameter λ vs. classification accuracy. Right: the perturbation angle vs. classification accuracy.

CIFAR-10

Method	M=200	M=500	M=2000
GEM (Lopez-Paz & Ranzato, 2017)	29.99±3.92	29.45±5.64	27.20±4.50
GSS (Aljundi et al., 2019b)	38.62±3.59	48.97±3.25	60.40±4.92
iCaRL (Rebuffi et al., 2017)	32.44±0.93	34.95±1.23	33.57±1.65
ER (Riemer et al., 2018)	49.07±1.65	61.58±1.12	76.89±0.99
ER w/ Gaussian	61.52±1.42	68.54±2.01	78.27±0.52
ER w/ WAP	**63.12±2.15**	**72.07±1.37**	**80.38±0.95**
DER++ (Buzzega et al., 2020)	64.88±1.17	72.70±1.36	78.54±0.97
DER++ w/ Gaussian	63.02±0.53	71.04±0.72	79.22±0.42
DER++ w/ WAP	**65.12±0.77**	**75.01±0.24**	**81.54±0.12**
ER-ACE (Caccia et al., 2021)	63.18±0.56	71.98±1.30	80.01±0.76
ER-ACE w/ Gaussian	65.21±0.89	72.01±0.76	78.92±0.58
ER-ACE w/ WAP	**66.56±0.81**	**72.86±1.02**	**80.24±0.50**

CIFAR-100

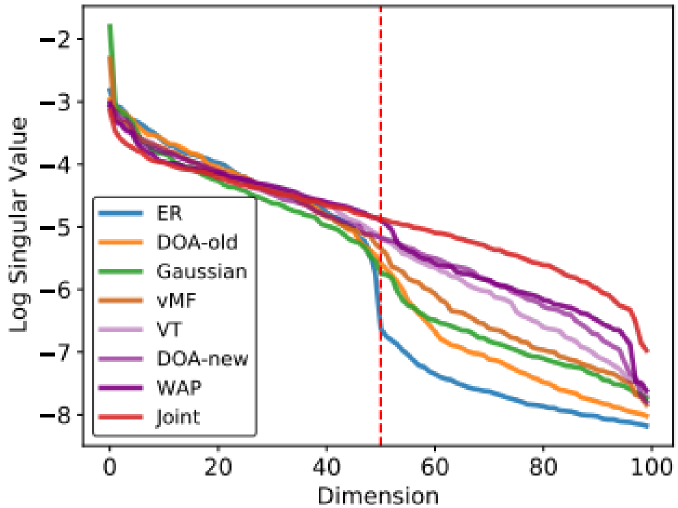
Method	M=200	M=500	M=2000
GEM (Lopez-Paz & Ranzato, 2017)	20.75±0.66	25.54±0.65	37.56±0.87
GSS (Aljundi et al., 2019b)	19.42±0.29	21.92±0.34	27.07±0.25
iCaRL (Rebuffi et al., 2017)	28.00±0.91	33.25±1.25	42.19±2.42
ER (Riemer et al., 2018)	22.14±0.42	31.02±0.79	43.54±0.59
ER w/ Gaussian	27.51±0.93	37.54±0.71	49.61±1.01
ER w/ WAP	**30.16±1.02**	**40.24±0.78**	**52.92±0.03**
DER++ (Buzzega et al., 2020)	29.68±1.38	39.08±1.76	54.38±0.86
DER++ w/ Gaussian	30.59±0.40	40.52±0.29	53.7±0.42
DER++ w/ WAP	**32.18±0.67**	**43.78±0.89**	**55.04±0.81**
ER-ACE (Caccia et al., 2021)	35.09±0.92	43.12±0.85	53.88±0.42
ER-ACE w/ Gaussian	37.01±0.70	44.57±0.83	54.84±0.12
ER-ACE w/ WAP	**37.46±0.77**	**45.79±0.73**	**56.02±0.64**

TinyImageNet

Method	M=200	M=500	M=2000
GEM (Lopez-Paz & Ranzato, 2017)	-	-	-
GSS (Aljundi et al., 2019b)	8.57±0.13	9.63±0.14	11.94±0.17
iCaRL (Rebuffi et al., 2017)	5.50±0.52	11.00±0.55	18.10±1.13
ER (Riemer et al., 2018)	8.65±0.16	10.05±0.28	18.19±0.47
ER w/ Gaussian	9.42±0.12	12.94±0.52	21.43±0.78
ER w/ WAP	**10.41±0.37**	**16.27±0.25**	**22.62±0.10**
DER++ (Buzzega et al., 2020)	10.96±1.17	19.38±1.41	**30.11±0.57**
DER++ w/ Gaussian	10.52±0.12	15.75±0.35	25.28±0.30
DER++ w/ WAP	**12.07±0.35**	**21.24±0.47**	29.33±0.71
ER-ACE (Caccia et al., 2021)	14.29±0.74	20.87±0.69	30.10±0.92
ER-ACE w/ Gaussian	16.72±0.41	22.82±0.39	30.92±0.41
ER-ACE w/ WAP	**17.05±0.22**	**23.56±0.85**	**32.54±0.72**

Ablation

- MOCA serves as a representation augmentation method:
- MOCA do improve the gradient diversity and approach the gradient under Joint training.
- Variation Towards New-Class is Important for Continual Learning:



Method	Perturbed	Original	Accuracy
Baseline	-	72.51	29.94
Minus New Feature	90.12	70.91	27.35
Add New Feature	71.34	**77.58**	**32.60**

Table 6: Adding perturbations in different directions: Towards the new-class feature or opposite to the new-class feature.

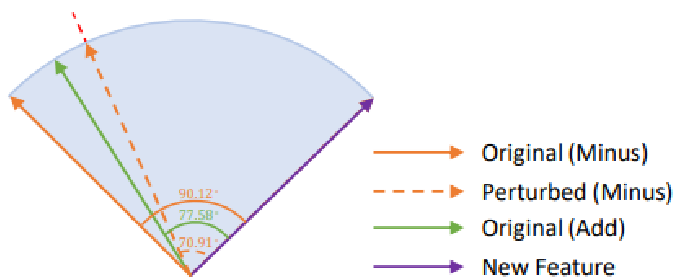


Figure 13: Different changes of the angle between old-class and new-class features by diversifying the feature towards or opposite the new-class manifold.