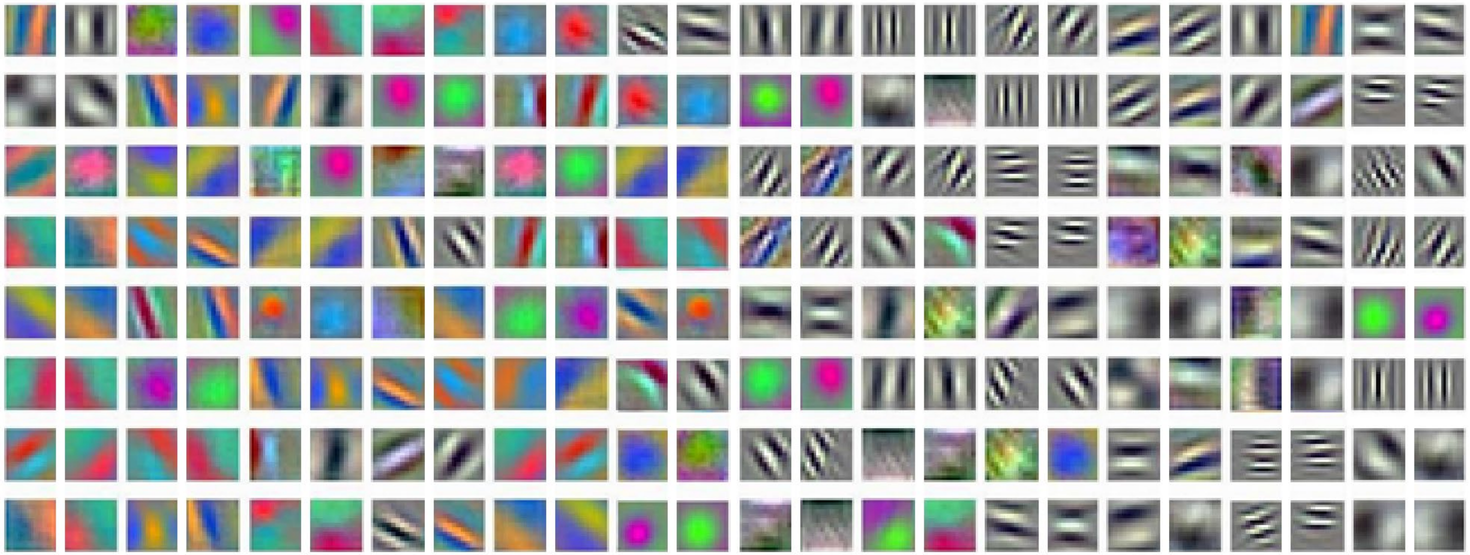
Learning towards Minimum Hyperspherical Energy

Published in NIPS 2018

Motivation

Filters learned in convolutional neural networks are usually highly redundant.

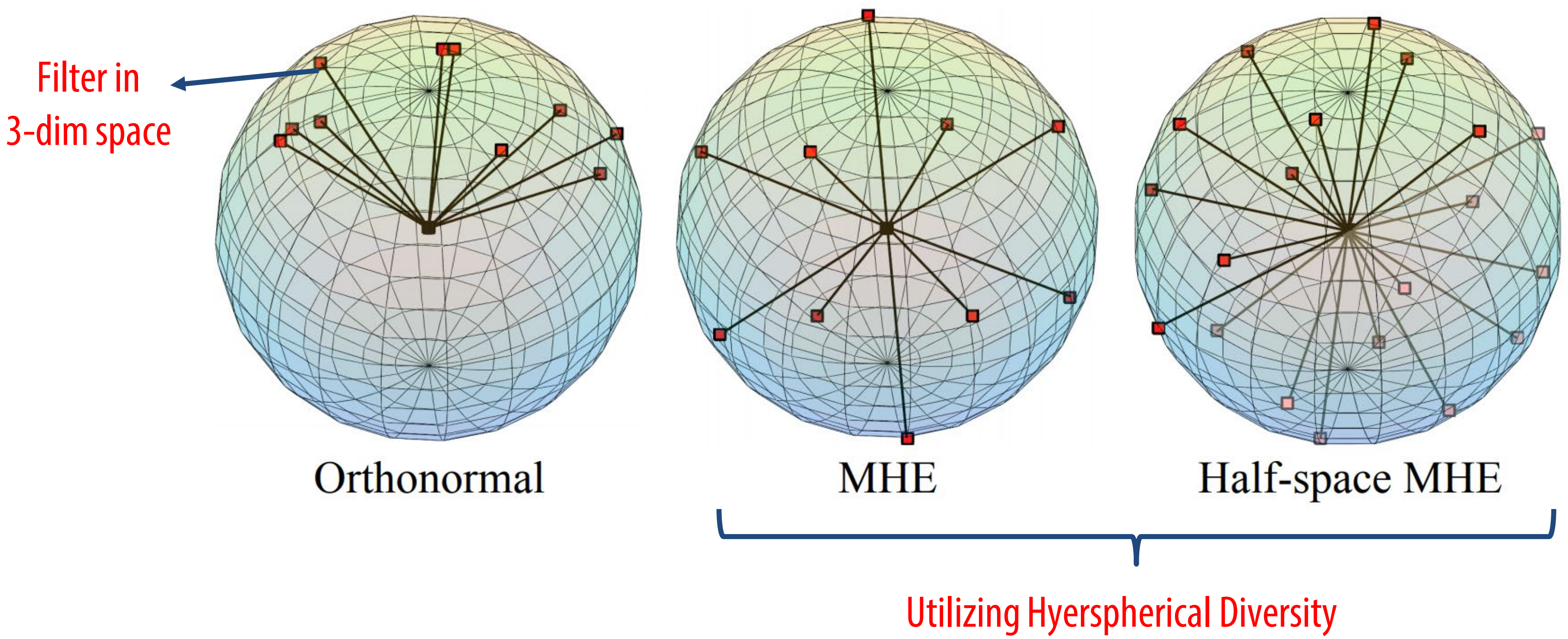
- Visualization of Conv1 filters from AlexNet



- We can observe that these filters are highly redundant and correlated.
- **Is there a good regularization to prevent the filters to be redundant?**

Intuition

- To avoid the redundancy, we need to first define a way to characterize diversity. The most straightforward way is to use orthogonality.
- However, orthogonality may still result in redundancy when the filter dimension is smaller than the number of filters.
- To better characterize diversity, we propose the **hyperspherical diversity** which can effectively reduce the redundancy and improve the network generalization.



Learning towards Minimum Hyperspherical Energy (MHE)

Hyperspherical Energy is defined to characterize the diversity on a hypersphere.

- We first define the hyperspherical energy functional for N neurons with (d+1)-dimension $\boldsymbol{W}_N = \{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_N \in \mathbb{R}^{d+1}\}$ as

$$\boldsymbol{E}_{s,d}(\hat{\boldsymbol{w}}_i|_{i=1}^N) = \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} f_s\left(\|\hat{\boldsymbol{w}}_i - \hat{\boldsymbol{w}}_j\|\right) = \begin{cases} \sum_{i\neq j} \|\hat{\boldsymbol{w}}_i - \hat{\boldsymbol{w}}_j\|^{-s}, & s>0 \\ \sum_{i\neq j} \log\left(\|\hat{\boldsymbol{w}}_i - \hat{\boldsymbol{w}}_j\|^{-1}\right), & s=0 \end{cases}$$

- where $f_s(\cdot)$ is a decreasing real-valued function, and $\hat{\boldsymbol{w}}_i = \frac{\boldsymbol{w}_i}{\|\boldsymbol{w}_i\|}$ is the i-th neuron weight projected onto the unit hypersphere.
- In this paper, we use as Riesz s-kernel:

$$f_s(z) = z^{-s}, s > 0$$

$$f_0(z) = \log(z^{-1})$$

- In fact, minimizing E_0 can also be viewed as a relaxation of minimizing E_s for s>0. (See our paper for more details.)

Variants of MHE

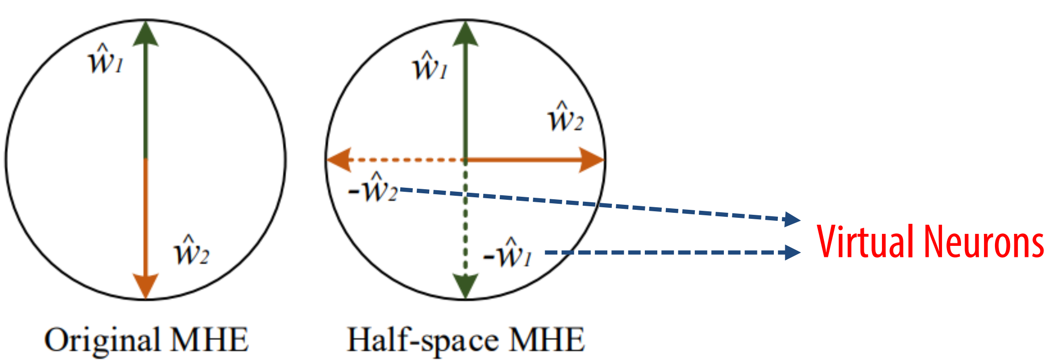
MHE beyond Euclidean Distance

- The hyperspherical energy is originally defined based on the Euclidean distance on a hypersphere, which can be viewed as an angular measure.
- In addition to Euclidean distance, we consider the geodesic distance (i.e., angle) on a unit hypersphere as a distance measure for neurons.

$$\boldsymbol{E}^{a}_{s,d}(\hat{\boldsymbol{w}}_i|_{i=1}^{N}) = \sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} f_s\big(\arccos(\hat{\boldsymbol{w}}_i^\top \hat{\boldsymbol{w}}_j)\big) = \begin{cases} \sum_{i\neq j} \arccos(\hat{\boldsymbol{w}}_i^\top \hat{\boldsymbol{w}}_j)^{-s}, & s>0 \\ \sum_{i\neq j} \log\big(\arccos(\hat{\boldsymbol{w}}_i^\top \hat{\boldsymbol{w}}_j)^{-1}\big), & s=0 \end{cases}$$

MHE in Half Space

- To avoid the collinear redundancy, we propose the half-space MHE.



Understanding MHE from decoupled view

- Inspired by decoupled networks [Liu et al. Decoupled Networks, CVPR 2018], we can view the original convolution as the multiplication of the angular function *g* and the magnitude function *h*:

$$f(\boldsymbol{w},\boldsymbol{x}) = h(\|\boldsymbol{w}\|, \|\boldsymbol{x}\|) \cdot g(\theta) = \big(\|\boldsymbol{w}\| \cdot \|\boldsymbol{x}\|\big) \cdot \big(\cos(\theta)\big)$$

- By combining MHE to a standard neural networks (e.g., CNNs), the entire regularization term becomes

$$\mathcal{L}_{\text{reg}} = \underbrace{\lambda_{\text{w}} \cdot \frac{1}{\sum_{j=1}^{L} N_j} \sum_{j=1}^{L} \sum_{i=1}^{N_j} \|\boldsymbol{w}_i\|}_{\text{Weight decay: regularizing the magnitude of kernels}} \underbrace{+ \lambda_{\text{h}} \cdot \sum_{j=1}^{L-1} \frac{1}{N_j(N_j-1)} \{\boldsymbol{E}_s\}_j + \lambda_{\text{o}} \cdot \frac{1}{N_L(N_L-1)} \boldsymbol{E}_s(\hat{\boldsymbol{w}}_i^{\text{out}}|_{i=1}^{c})}_{\text{MHE: regularizing the direction of kernels}}$$

- From the decoupled view, we can see that MHE is actually very meaningful in regularizing the neural networks, and it also serves as a complementary role to weight decay.

Ablation Study I

- Variants of MHE.

Method	CIFAR-10			CIFAR-100		
	$s=2$	$s=1$	$s=0$	$s=2$	$s=1$	$s=0$
MHE	6.22	6.74	6.44	27.15	27.09	**26.16**
Half-space MHE	6.28	6.54	**6.30**	**25.61**	**26.30**	26.18
A-MHE	**6.21**	6.77	6.45	26.17	27.31	27.90
Half-space A-MHE	6.52	**6.49**	6.44	26.03	26.52	26.47
Baseline	7.75			28.13		

Table 1: Testing error (%) of different MHE on CIFAR-10/100.

- Network width.

Method	16/32/64	32/64/128	64/128/256	128/256/512	256/512/1024
Baseline	47.72	38.64	28.13	24.95	25.45
MHE	36.84	30.05	26.75	24.05	23.14
Half-space MHE	**35.16**	**29.33**	**25.96**	**23.38**	**21.83**

Table 2: Testing error (%) of different width on CIFAR-100.

- Network depth.

Method	CNN-6	CNN-9	CNN-15
Baseline	32.08	28.13	N/C
MHE	28.16	26.75	26.9
Half-space MHE	**27.56**	**25.96**	**25.84**

Table 3: Testing error (%) of different depth on CIFAR-100. N/C: not converged.

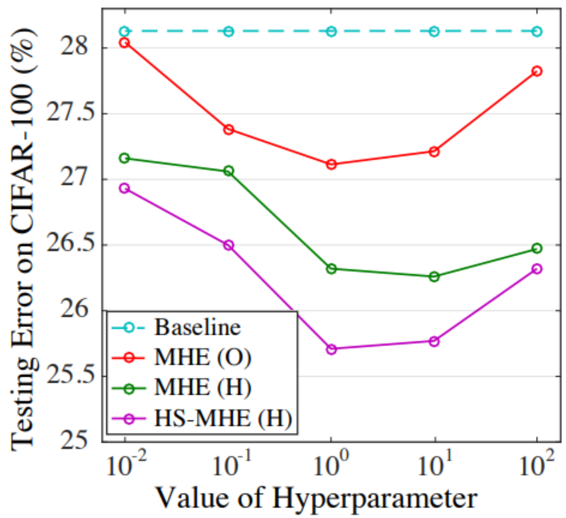
Ablation Study II

- MHE for regularizing hidden layers (H), output layers (O), or both.

Method	H O × √	H O √ ×	H O √ √
MHE	26.85	26.55	**26.16**
Half-space MHE	N/A	26.28	**25.61**
A-MHE	27.8	26.56	**26.17**
Half-space A-MHE	N/A	26.64	**26.03**
Baseline		28.13	

Table 4: Ablation study on CIFAR-100.

- Hyperparameter experiment. It shows that MHE is not sensitive to the selection of hyperparameters.



MHE for Image Classification

- ResNet-32 with MHE for CIFAR-10 and CIFAR-100

Method	CIFAR-10	CIFAR-100
ResNet-110-original [15]	6.61	25.16
ResNet-1001 [16]	4.92	**22.71**
ResNet-1001 (64 batch) [16]	**4.64**	-
baseline	5.19	22.87
MHE	4.72	22.19
Half-space MHE	**4.66**	**22.04**

Table 5: Error (%) of ResNet-32.

- Large-scale Object Recognition on ImageNet-2012

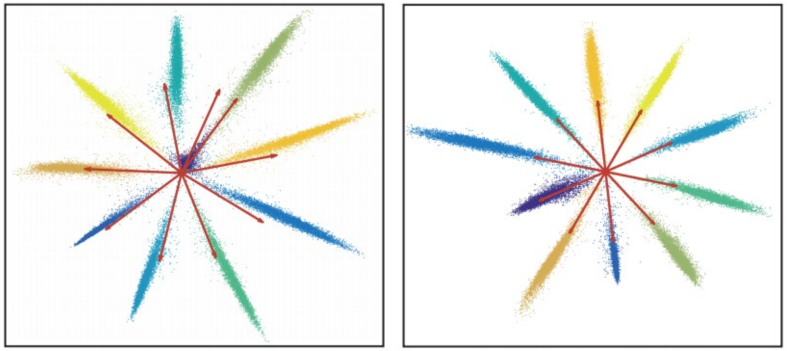
Method	ResNet-18	ResNet-34
baseline	33.95	30.04
Orthogonal [37]	33.65	29.74
Orthnormal	33.61	29.75
MHE	33.50	29.60
Half-space MHE	**33.45**	**29.50**

Table 6: Top1 error (%) on ImageNet.

We can observe that MHE and half-space MHE can consistently improve the classification accuracy by a significant margin.

MHE for Class-imbalance Learning

- MHE can alleviate the class-imbalance problem, and therefore improve the accuracy on class-imbalance learning.
- 2D feature visualization on MNIST



(a) CNN without MHE (b) CNN with MHE

We can observe that CNN w/ MHE can learn reasonable feature distribution even if the training dataset is highly imbalanced, while CNN w/o MHE can not.

MHE for Face Recognition

- We apply MHE to the loss function of Sphereface [Liu et al. SphereFace: Deep Hypersphere Embedding for Face Recognition, CVPR 2017], and propose SphereFace+ with the following loss function:

$$\mathcal{L}_{\text{SF+}} = \underbrace{\frac{1}{m}\sum_{j=1}^{m}\ell_{\text{SF}}(\langle \boldsymbol{w}_i^{\text{out}}, \boldsymbol{x}_j\rangle_{i=1}^{c}, \boldsymbol{y}_j, m_{\text{SF}})}_{\text{Angular softmax loss: promoting } \textbf{intra-class compactness}} + \underbrace{\lambda_{\text{M}} \cdot \frac{1}{m(N-1)}\sum_{i=1}^{m}\sum_{j=1, j\neq y_i}^{N} f_s(\|\hat{\boldsymbol{w}}_{y_i}^{\text{out}} - \hat{\boldsymbol{w}}_j^{\text{out}}\|)}_{\text{MHE: promoting } \textbf{inter-class separability}}$$

- Performance comparison to the state-of-the-art

Method	LFW	MegaFace
Softmax Loss	97.88	54.86
Softmax+Contrastive [46]	98.78	65.22
Triplet Loss [41]	98.70	64.80
L-Softmax Loss [30]	99.10	67.13
Softmax+Center Loss [55]	99.05	65.49
CosineFace [53, 51]	**99.10**	**75.10**
SphereFace	99.42	72.72
SphereFace+ (ours)	**99.47**	**73.03**

Improving GANs with MHE

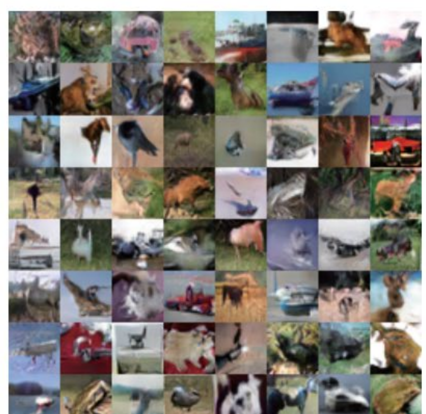
- Combining MHE to the discriminator of GANs can significantly improve the generation quality:

Method	Inception score
Real data	11.24±.12
Weight clipping	6.41±.11
GAN-gradient penalty (GP)	6.93±.08
WGAN-GP [9]	6.68±.06
Batch Normalization [21]	6.27±.10
Layer Normalization [2]	7.19±.12
Weight Normalization [40]	6.84±.07
Orthonormal [4]	7.40±.12
SN-GANs [35]	**7.42**±.08
MHE (ours)	7.32±.10
MHE + SN [35] (ours)	**7.59**±.08

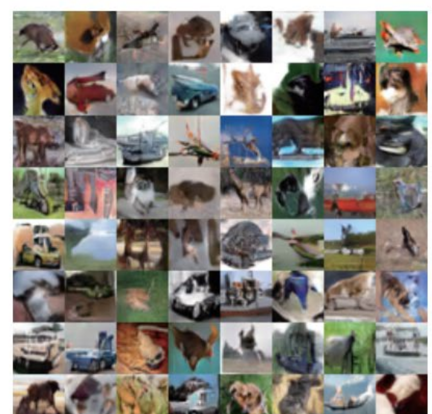
Table 14: Inception scores with unsupervised image generation on CIFAR-10.

- Sample images:

Dataset



Baseline GAN



GAN with MHE and SN

Thank you!