# Learning towards Minimum Hyperspherical Energy

Weiyang Liu[1]*, Rongmei Lin[2]*, Zhen Liu[1]*, Lixin Liu[1], Zhiding Yu[3], Bo Dai[4], Le Song[1]

1. Georgia Tech 2. Emory University 3. NVIDIA 4. Google Brain * Equal Contribution

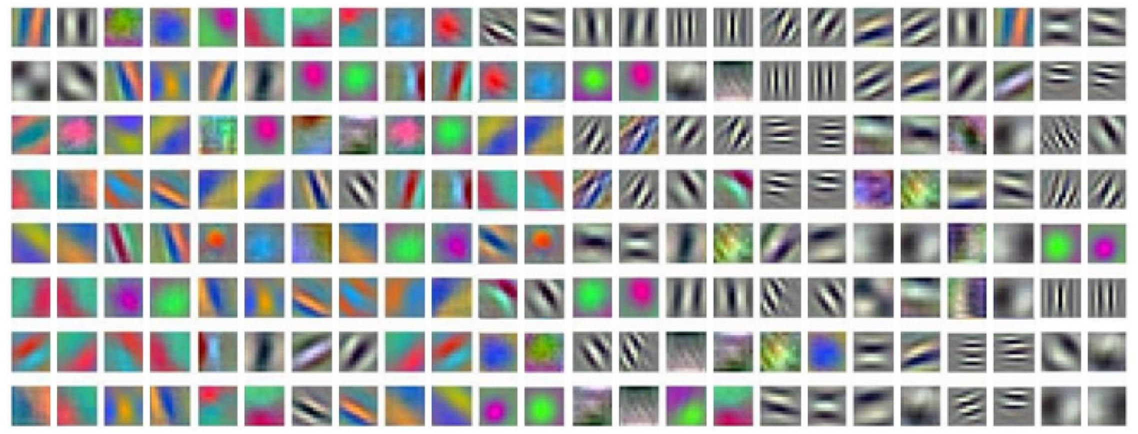
Code for SphereFace+

Code for MHE

## Introduction

### Motivation

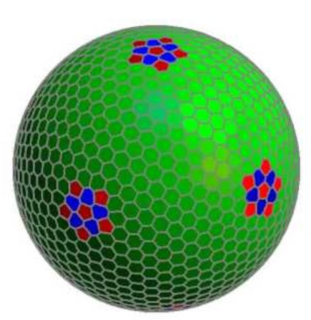
- Filters learned in convolutional neural networks are highly redundant. (e.g. Conv1 filters from AlexNet)



- Recent studies show that reducing the neuron redundancy can effectively improve the network generalization.
- A natural way is to use orthogonality, but it may not be effective when the filter dimension is smaller than the number of filters.

### Connection to physics

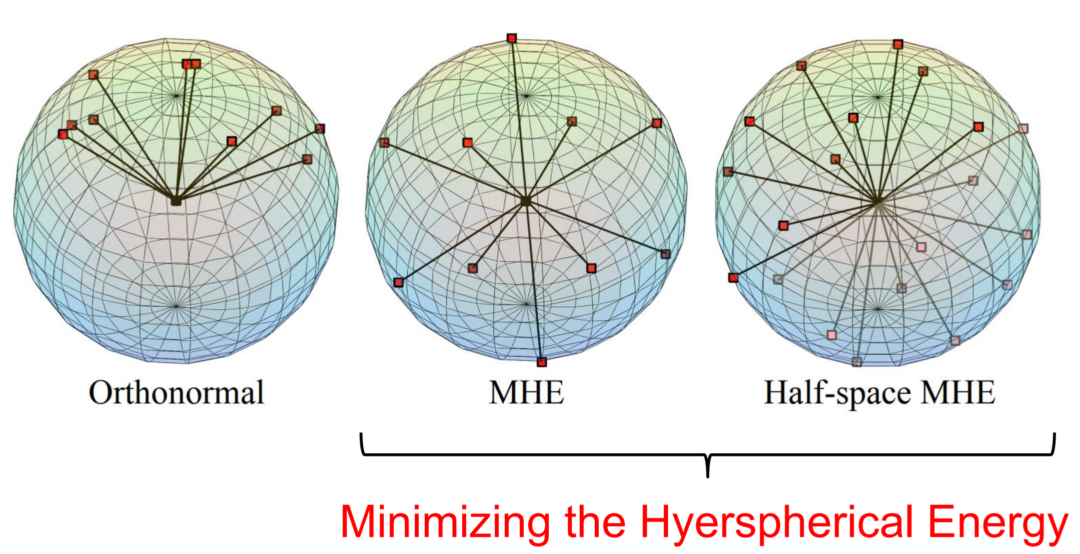
- To characterize diversity, we draw inspiration from a famous physics problem, called **Thomson problem**.
- Thomson problem is to find a state that distributes N electrons on a unit sphere as evenly as possible to minimize potential energy.



- The electrons repel each other with a force given by Coulomb's Law.

### Intuition

- We draw inspiration from Thomson problem, and propose **hyperspherical energy** to characterize neuron diversity.
- The intuitive comparison is shown as follows:




## Minimum Hyperspherical Energy (MHE)

- Hyperspherical Energy characterizes the diversity of neurons on a hypersphere.
- We 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

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

where

$f_s(\cdot)$ = Riesz s-kernel, with $f_s(z)=z^{-s}, s>0$, $f_0(z)=\log(z^{-1})$

$\hat{\boldsymbol{w}}_i=\frac{\boldsymbol{w}_i}{\|\boldsymbol{w}_i\|}$ = normalized weight of the *i*-th neuron

- In fact, $f_s(\cdot)$ can be a general decreasing function.
- Minimizing $E_0$ can be viewed as a relaxation of minimizing $E_s$ for s>0.
- We add this energy to the total regularization loss in a network and minimize it via SGD and back-prop.

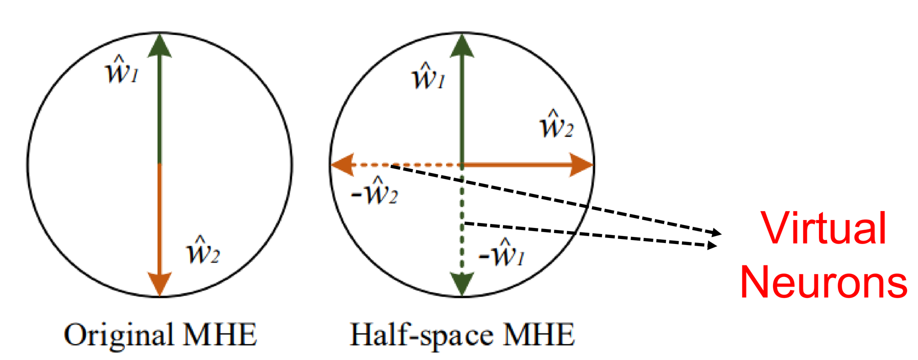
### MHE beyond Euclidean Distance

- In addition to Euclidean distance, we consider the geodesic distance (i.e., angle) on a unit hypersphere as a distance measure for neurons.
- The formulation is given as follows:

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

### MHE in Half Space

- The original MHE suffers from **collinear redundancy**, as shown in the following :




- Instead, we can construct virtual neurons in the opposite directions of the original neurons.
- We minimize the half-space hyperspherical energy of both original and virtual neurons together to encourage a diverse distribution of them.

## Theoretical Properties

- The optimal distribution of N neurons (w.r.t. MHE) asymptotically converge to the **uniform distribution on the hypersphere** as N becomes larger.
- Minimizing MHE can **provably guarantee generalization** error in a one-hidden-layer net under some assumptions.

## Decoupled View of MHE

- We can decouple the convolutional into magnitude and angle [Liu et al. Decoupled Networks, CVPR 2018]

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

- MHE is complementary to weight decay:

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

## Ablation Study and Experiments

- **Evaluation of different variants of MHE**.
  - A-MHE = MHE with angular distance.
  - Different s represents using different energy formulation.

| 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.

- **Different 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.

- **Different 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.

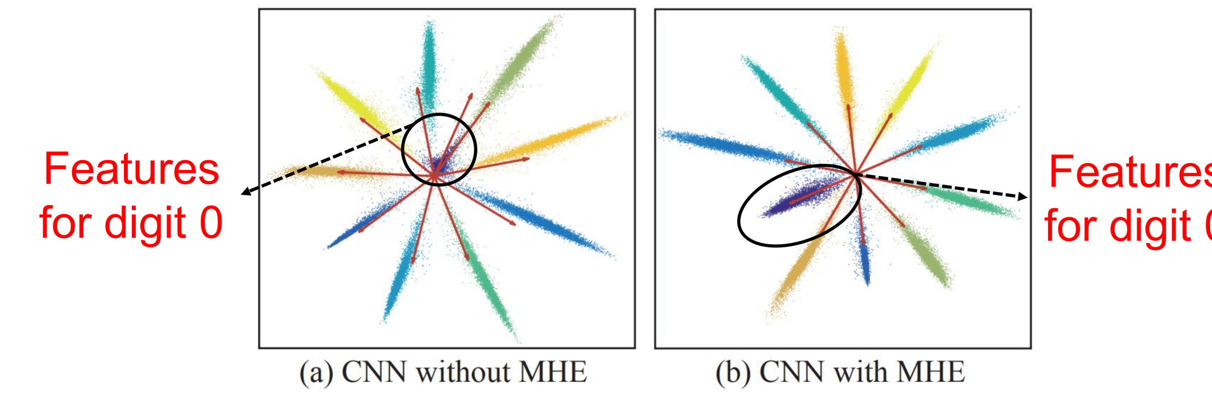
## ImageNet Classification

- MHE can effectively improve the accuracy of existing networks on image recognition.

| 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** |

## Class-imbalance Learning

- We first randomly throw away 98% training data for digit 0 in MNIST (only 100 samples are preserved for digit 0), and then train a 6-layer CNN on this imbalance MNIST. The 2D features are visualized as follows (Red arrows denote the classifier neurons):




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

- When MHE is applied to the output layers, MHE can greatly alleviate the class imbalance problem in the training set and help to learn reasonable features.

## SphereFace+: MHE for Face Recognition

- SphereFace is a state-of-the-art face recognition method.
- SphereFace+ applies MHE regularization to the output layer in addition to the loss function of SphereFace.

| $m_{\text{SF}}$ | LFW | | MegaFace | |
|---|---|---|---|---|
| | SphereFace | SphereFace+ | SphereFace | SphereFace+ |
| 1 | 96.35 | **97.15** | 39.12 | **45.90** |
| 2 | 98.87 | **99.05** | 60.48 | **68.51** |
| 3 | 98.97 | **99.13** | 63.71 | **66.89** |
| 4 | 99.26 | **99.32** | 70.68 | **71.30** |

Performance on 20-layer ResNet

| $m_{\text{SF}}$ | LFW | | MegaFace | |
|---|---|---|---|---|
| | SphereFace | SphereFace+ | SphereFace | SphereFace+ |
| 1 | 96.93 | **97.47** | 41.07 | **45.55** |
| 2 | 99.03 | **99.22** | 62.01 | **67.07** |
| 3 | 99.25 | **99.35** | 69.69 | **70.89** |
| 4 | 99.42 | **99.47** | 72.72 | **73.03** |

Performance on 64-layer ResNet

- SphereFace+ consistently outperforms SphereFace, showing that MHE can improve generalization.

| 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** |

Comparison to the state-of-the-art

## MHE for GANs

- **MHE can also be applied to improve the image generation of GANs**, and is complementary to spectral normalization. See our paper for details.