Learning towards Minimum Hyperspherical Energy

Weiyang Liu1*, Rongmei Lin2*, Zhen Liu1*, Lixin Liu1, Zhiding Yu3, Bo Dai4, Le Song1

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 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 \( \mathbf{w}_i = [w_{i1}, \ldots, w_{id}, y_i] \) as

\[
E_{MHE}(\mathbf{w}_1, \ldots, \mathbf{w}_N) = \sum_{i=1}^{N} \sum_{j\neq i}^{N} \frac{f_s(\mathbf{w}_i^T \mathbf{w}_j)}{\|\mathbf{w}_i\| \|\mathbf{w}_j\|} = \sum_{i=1}^{N} \sum_{j\neq i}^{N} \frac{f_s(\|\mathbf{w}_i\| \|\mathbf{w}_j\| \mathbf{w}_i^T \mathbf{w}_j)}{\|\mathbf{w}_i\| \|\mathbf{w}_j\|} + \frac{\sum_{i=1}^{N} f_s(\|\mathbf{w}_i\|)}{N} = \sum_{i=1}^{N} \sum_{j\neq i}^{N} \frac{f_s(\|\mathbf{w}_i\| \|\mathbf{w}_j\| \mathbf{w}_i^T \mathbf{w}_j)}{\|\mathbf{w}_i\| \|\mathbf{w}_j\|}, \quad s > 0
\]

where

\[
f_s(z) = \frac{z^s}{z^s - 1}, \quad f_s(z) = \frac{z^s}{z^s - 1}, \quad s > 0
\]

- In fact, \( f_s(z) \) can be a general decreasing function.
- 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:

MHE in Half Space

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

![Collinear Neurons](image1.png)

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 large.
- 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)
- Different s represents using different energy formulation.
- A-MHE = MHE with angular distance.

Ablation Study and Experiments

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

<table>
<thead>
<tr>
<th>Method</th>
<th>LPI</th>
<th>MeanFace</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet</td>
<td>26.91</td>
<td>26.79</td>
</tr>
<tr>
<td>SphereFace+ MHE</td>
<td>26.31</td>
<td>26.14</td>
</tr>
<tr>
<td>SphereFace</td>
<td>26.35</td>
<td>26.17</td>
</tr>
</tbody>
</table>

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

Different network depth:

<table>
<thead>
<tr>
<th>Method</th>
<th>LPI</th>
<th>MeanFace</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet</td>
<td>26.91</td>
<td>26.79</td>
</tr>
<tr>
<td>SphereFace+ MHE</td>
<td>26.31</td>
<td>26.14</td>
</tr>
<tr>
<td>SphereFace</td>
<td>26.35</td>
<td>26.17</td>
</tr>
</tbody>
</table>

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

Different network width:

<table>
<thead>
<tr>
<th>Method</th>
<th>LPI</th>
<th>MeanFace</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet</td>
<td>26.91</td>
<td>26.79</td>
</tr>
<tr>
<td>SphereFace+ MHE</td>
<td>26.31</td>
<td>26.14</td>
</tr>
<tr>
<td>SphereFace</td>
<td>26.35</td>
<td>26.17</td>
</tr>
</tbody>
</table>

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

ImageNet Classification

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

Class-imbalance Learning

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

- 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.

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.

Code for SphereFace+

- Code for MHE

Code for SphereFace

Performance on 64-layer ResNet

<table>
<thead>
<tr>
<th>Method</th>
<th>LPI</th>
<th>MeanFace</th>
</tr>
</thead>
<tbody>
<tr>
<td>SphereFace+ MHE</td>
<td>73.95</td>
<td>73.70</td>
</tr>
<tr>
<td>SphereFace</td>
<td>73.88</td>
<td>73.69</td>
</tr>
</tbody>
</table>

Performance on 6-layer ResNet

<table>
<thead>
<tr>
<th>Method</th>
<th>LPI</th>
<th>MeanFace</th>
</tr>
</thead>
<tbody>
<tr>
<td>SphereFace+ MHE</td>
<td>73.95</td>
<td>73.70</td>
</tr>
<tr>
<td>SphereFace</td>
<td>73.88</td>
<td>73.69</td>
</tr>
</tbody>
</table>

Table 4: Testing error (%) of different MHE on SphereFace.

Comparison to the state-of-the-art