Nonlocal Neural Networks, Nonlocal Diffusion and Nonlocal Modeling
Yunzhe Tao¹, Qi Sun², Qiang Du¹, and Wei Liu³
¹: Columbia University; ²: CSRC; ³: Tencent AI Lab

Introduction

We study the nature of diffusion and damping effect of nonlocal networks by doing spectrum analysis on the well-trained networks, and then propose a new formulation of the nonlocal block from nonlocal modeling perspective, which is more robust.

Damping Effect of Nonlocal Net

Nonlocal block in [Wang et al., 2018]:
- Weight matrix
- Affinity function

\[ Z^{k+1} = Z^k + \frac{W^k}{C_k(Z^k)} \sum \omega(Z^k, Z^k) g(Z^k) \]

Incorporating nonlocal blocks into ResNet:

\[ Z^{k+1} = Z^k + \mathcal{F}(Z^k; W^k) \]

Training ResNet-20 with different number of nonlocal blocks on CIFAR-10:
- Works well when adding 1-3 nonlocal blocks
- Training becomes difficult when adding 4 blocks
- The original nonlocal network is not robust to multiple nonlocal blocks

Doing spectrum analysis on symmetrized \( W^k \) when \( k \)-th block is a nonlocal block:
- Computing the eigenvalues of
  \[ \mathcal{W}^k = \frac{W^k W^k + (W^k)^T W^k}{2} \]
- The eigenvalues are all real numbers
- The eigenvalues with greatest magnitudes describe the characteristics of the weights in nonlocal blocks

Eigenvectors of adding 2 nonlocal blocks:
- Nonlocal block or residual block

Eigenvectors of adding 4 nonlocal blocks:
- Nonlocal block or residual block

New Formulation of Nonlocal Net

New nonlocal stage (one or more blocks):

\[ Z_i^{n+1} = Z_i^n + \frac{W^i}{C_i(x)} \sum \omega(x_i, x_i)(Z_i^n - Z_i^n) \]

- Eigenvalues of adding 4 nonlocal blocks

Validation errors of different models:
(baseline: ResNet-20)

Connection to Nonlocal Modeling

- Nonlocal diffusion process
- The discrete nonlocal diffusion:
  \[ Z^{n+1} = Z^n + \sum \omega(x_i, x_i)(Z_i^n - Z_i^n) \]

Continuum form:
\[
\begin{cases}
    x_t(x, t) - \mathcal{L}x(x) = 0 \\
    x(x, 0) = u(x)
\end{cases}
\]

where
\[ \mathcal{L}x(x) = \int \rho(x, y)(x(y) - x(x))dy \]

is a Hilbert-Schmidt operator
- Stable in finite time for both time directions (for both positive and negative weights)
- Markov jump process
- Time-homogeneous Markov jump process:

\[ x(x, t) = \int p(y, x)x(y, t)dy \]

Discrete-time Markov jump process:
\[ Z_i^{n+1} = Z_i^n - \sum_j \mathcal{K}_{ij}(Z_j^n - Z_i^n) \]

\( \mathcal{K} \) is a Markov matrix, thus is Hilbert-Schmidt
- Stable in finite time for both forward and backward jump process

Table 1: Validation errors of different models based on PreResNet-20 over CIFAR-10.

<table>
<thead>
<tr>
<th>Model</th>
<th>Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>8.19</td>
</tr>
<tr>
<td>2-block (original)</td>
<td>7.83</td>
</tr>
<tr>
<td>3-block (original)</td>
<td>8.28</td>
</tr>
<tr>
<td>4-block (original)</td>
<td>15.02</td>
</tr>
<tr>
<td>The Same Place</td>
<td></td>
</tr>
<tr>
<td>2-block (proposed)</td>
<td>7.74</td>
</tr>
<tr>
<td>3-block (proposed)</td>
<td>7.62</td>
</tr>
<tr>
<td>4-block (proposed)</td>
<td>7.37</td>
</tr>
<tr>
<td>5-block (proposed)</td>
<td>7.29</td>
</tr>
<tr>
<td>6-block (proposed)</td>
<td>7.55</td>
</tr>
<tr>
<td>Different Places</td>
<td></td>
</tr>
<tr>
<td>3-block (original)</td>
<td>8.07</td>
</tr>
<tr>
<td>3-block (proposed)</td>
<td>7.33</td>
</tr>
</tbody>
</table>