INTRODUCTION

CNNs can suffer from diverse issues, such as:
- Exploding, vanishing gradients
- Scaling-based weight space symmetry
- Covariant-shift

We propose a weight soft-regularization method based on the Oblique manifold.

PRELIMINARIES

Given weight vector:
\[ W \in \mathbb{R}^{n \times p}, \]
\( p \) is the dimensionality of each weight vector of a filter, while \( n \) is the number of filters.

The Oblique manifold defines:
\[ \text{diag}(WW^T) = I \]

CONNECTION TO WD

Weight decay update rule:
\[ \frac{\partial L_{\text{wd}}}{\partial w_{ijc, co}} = 2w_{ijc, co} \]

Norm Loss update rule:
\[ \frac{\partial L_{\text{nl}}}{\partial w_{ijc, co}} = 2w_{ijc, co} \left( 1 - \frac{1}{\| w_{co} \|} \right) \]

The Norm Loss can be seen as an extension of the weight decay where the weight decay factor and its sign are regulated during training by the norm of the weight vector.

RELATED WORK

Weight regularization
- Weight decay
- Weight normalization
- Weight orthogonalization

Activation normalization
- Batch normalization
- Group normalization
- Kalman normalization

PROPOSED METHOD, THE NORM LOSS (NL)

\[
L_{\text{nl}} = \sum_{c_o=1}^{C_o} \left( 1 - \frac{1}{\| w_{co} \|} \right) F_w \sum_{c_i=1}^{C_i} \sum_{j=1}^{C_{ij}} w_{ijc_i, co}^2
\]

where \( F_w, F_h, C_i, C_o \) are the filter (or weight vector) width, height, number of input and number of output channels respectively. The loss is penalizing the weight vector of each neuron if its Euclidean norm is different from one. The final loss function then becomes:
\[
L_{\text{total}} = L_{\text{target}} + \lambda_{nl} \cdot L_{\text{nl}}
\]

Effect: Slowly steer the weight vectors to unit norm.

CROSS ENTROPY EVOLUTION (CIFAR-10)

<table>
<thead>
<tr>
<th>Top-1 error</th>
<th>Top-5 error</th>
</tr>
</thead>
<tbody>
<tr>
<td>wd (repr)</td>
<td>25.29</td>
</tr>
<tr>
<td>nl (Ours)</td>
<td>24.34</td>
</tr>
<tr>
<td>wd</td>
<td>23.85</td>
</tr>
<tr>
<td>ONI</td>
<td>23.30</td>
</tr>
</tbody>
</table>

VARYING BATCH SIZE

CIFAR-10

<table>
<thead>
<tr>
<th>Top-1 error</th>
<th>Top-5 error</th>
</tr>
</thead>
<tbody>
<tr>
<td>wd (repr)</td>
<td>3.73</td>
</tr>
<tr>
<td>wd</td>
<td>-</td>
</tr>
<tr>
<td>ONI</td>
<td>3.73</td>
</tr>
</tbody>
</table>

VARYING REG. FACTOR

<table>
<thead>
<tr>
<th>Top-1 error</th>
<th>Top-5 error</th>
</tr>
</thead>
<tbody>
<tr>
<td>wd (repr)</td>
<td>4.47 (4.662)</td>
</tr>
<tr>
<td>wd</td>
<td>-</td>
</tr>
<tr>
<td>ONI</td>
<td>4.47 (4.662)</td>
</tr>
</tbody>
</table>

CONCLUSIONS

Norm Loss:
- Comparable and sometimes better performance to the state of the art on popular architectures and benchmarks
- Lower computational complexity than most weight regularization methods
- High convergence speed
- Less sensitive to hyper parameters such as batch size and regularization factor