Beyond cross-entropy: learning highly separable feature distributions for robust and accurate classification

Arslan Ali, Andrea Migliorati, Tiziano Bianchi, Enrico Magli

Politecnico di Torino
GCCS - Motivation

- GCCS: Gaussian class-conditional simplex loss
- High Separability between classes
- Extract discriminative features from the input data
- Map the features to well behaved, target Gaussian distributions
**GCCS - Loss**

- **Assumption:** network output tends to a Gaussian distribution
  - compute batch statistics for each class
  - minimize KL divergence between target and obtained distribution

**Proposed loss**

Authorized users loss

\[
L_i = \frac{1}{i} \left[ \log \frac{\Sigma_{T_i}}{\Sigma_{O_i}} - D + \text{tr} \left( \Sigma_{T_i}^{-1} \Sigma_{O_i} \right) + (\mu_{T_i} - \mu_{O_i})^T \Sigma_{T_i}^{-1} (\mu_{T_i} - \mu_{O_i}) \right]
\]

\[
\mathcal{K}_i = \left( \frac{x - \mu_{O_i}}{\sigma_{O_i}} \right)^4
\]

Total loss:

\[
L^{GCCS} = \sum_{i=1}^{D} \left[ L_i + \lambda (\mathcal{K}_i - 3) \right]
\]
GCCS - Decision rule

• Partition the decision space into Voronoi regions

• Compute the distance from all the centers - choose minimum

\[ \hat{y} = \underset{i}{\arg \max} z_i, \]

• Index of the predicted class for the test image
GCCS - Advantages

• Equidistant classes
• Uniformity of feature distributions - lack of short path
• Higher robustness
• Simple straightforward decision boundaries
Results: Maximum accuracy

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GCCS - regular training</td>
<td>99.58</td>
<td>92.69</td>
<td>94.20</td>
<td>82.97</td>
<td>96.19</td>
<td>76.53</td>
</tr>
<tr>
<td>GCCS - fine tuning</td>
<td>99.64</td>
<td>93.83</td>
<td>95.58</td>
<td>81.52</td>
<td>97.06</td>
<td>77.48</td>
</tr>
<tr>
<td>No Defense - cross-entropy</td>
<td>99.35</td>
<td>91.91</td>
<td>94.12</td>
<td>78.59</td>
<td>95.78</td>
<td>76.30</td>
</tr>
<tr>
<td>Jacobian Reg. - regular training [60]</td>
<td>98.99</td>
<td>91.79</td>
<td>94.11</td>
<td>70.09</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Jacobian Reg. - fine-tuning[60]</td>
<td>98.53</td>
<td>92.43</td>
<td>93.54</td>
<td>82.09</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Input Gradient Reg. - regular training [53]</td>
<td>97.98</td>
<td>88.45</td>
<td>93.77</td>
<td>78.32</td>
<td>96.50</td>
<td>74.89</td>
</tr>
<tr>
<td>Input Gradient Reg. - fine-tuning [53]</td>
<td>99.11</td>
<td>92.55</td>
<td>93.17</td>
<td>76.15</td>
<td>96.90</td>
<td>75.68</td>
</tr>
<tr>
<td>Cross Lipschitz regular training [59]</td>
<td>96.78</td>
<td>92.54</td>
<td>91.42</td>
<td>80.10</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cross Lipschitz - fine-tuning [59]</td>
<td>98.77</td>
<td>92.41</td>
<td>93.50</td>
<td>79.39</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Tab. 2 Maximum test accuracy obtained through regular training vs fine-tuning over different benchmark datasets with different competing techniques in the case in which no adversarial attack is performed.

- GCCS yields high classification accuracy both for regular training and fine tuning
Adversarial attacks - Whitebox

- PGD
  - iterative FGSM
  - \( x_{t+1} = \text{Proj}\{x_t + \alpha \cdot \text{sign}[\nabla_x J(\theta, x_t, y)]\} \)

- TGSM
  - descending the gradient towards target class
  - \( x' = x - \epsilon \cdot \text{sign}[\nabla_x J(\theta, x, y')] \)

- JSMA
  - select the features to be altered to get desired output
  - \( \nabla l(x) = \frac{\partial l(x)}{\partial x} = \left[ \frac{\partial l(y)}{\partial x_j} \right]_{j=1,\ldots,M_{\text{out}}} \)
Robustness to targeted attacks

TGSM-5

Test accuracy when applying the TGSM attack (5 steps) for (a) ([MNIST, ResNet-18]) ; (b) ([SVHN, ResNet-18]); (c) ([CIFAR-10, ResNet-18]) (d) ([CIFAR-10, Shake-Shake-96]), for different values of $\epsilon$. 
Conclusions

• Novel loss promoting class separability and robustness

• High classification accuracy

• High robustness against adversarial attacks

arslan.ali@polito.it