**Motivation**

**Fact:** Deep CNNs tend to overfit the training data. Regularization is essential to prevent excessive co-adaptation of hidden units.

**Observation 1:** Low entropy output distributions are highly correlated with the L2-norm of the descriptor (penultimate layer).

**Observation 2:** High L2-norm descriptors are characterized by highly-valued spikes.

**Idea:** Regularize the training process by penalizing overconfident output distributions.

**Solution:** Drop-out a fraction of most active neurons (correlated with low-entropy distributions), proportionally to the predicted probability of the actual class.

**Code available:** https://github.com/clferrari/probability-guided-maxout

---

**Method**

1. **Forward pass to estimate the actual class probability.**
   \[
   \hat{y} = \frac{\exp(F(x))}{\sum_{j=1}^{C} \exp(F(x_j))}, \quad P_{gt} = y \cdot \hat{y}.
   \]

2. **Estimate percentage** \(\rho\) **of units to drop based on** \(P_{gt}\)
   \[
   \rho(x) = \gamma x; \quad x \in [0, 1], \quad \gamma = \frac{1}{2}
   \]

3. **Build the dropout mask** \(M \in [0, 1]^{d}\) **by sorting values of the descriptor** \(f \in \mathbb{R}^{d}\) **in descending order. Apply the permutation to** \(M\) **and drop-out a number** \(p = d \rho\) **units:**
   \[
   \hat{f} = f \circ M
   \]

4. **Scale the masked descriptor to maintain the expected output across train and inference with a learnable, per-sample scale factor** \(s = \frac{\alpha}{(1 - \rho)}\)

**N.B.** Steps 2-3 do not need gradient computation and are detached from the computational graph.

---

**Results**

**Ablation:** Fixed feature scaling

Finding the correct \(\alpha\) can be complex. Let the network learn its value.

**Test on benchmark datasets**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>Rank-1</th>
<th>Acc</th>
<th>Test Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIFAR10</td>
<td>CE</td>
<td>99.69</td>
<td>0.006</td>
<td></td>
</tr>
<tr>
<td>CIFAR10</td>
<td>Dropout</td>
<td>99.71</td>
<td>0.003</td>
<td></td>
</tr>
<tr>
<td>CIFAR10</td>
<td>PGM</td>
<td>99.73</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>CIFAR100</td>
<td>CE</td>
<td>68.47</td>
<td>2.07</td>
<td></td>
</tr>
<tr>
<td>CIFAR100</td>
<td>Dropout</td>
<td>69.65</td>
<td>2.21</td>
<td></td>
</tr>
<tr>
<td>CIFAR100</td>
<td>PGM</td>
<td>69.18</td>
<td>1.52</td>
<td></td>
</tr>
<tr>
<td>Caltech256</td>
<td>CE</td>
<td>62.21</td>
<td>2.28</td>
<td></td>
</tr>
<tr>
<td>Caltech256</td>
<td>Dropout</td>
<td>61.61</td>
<td>2.30</td>
<td></td>
</tr>
<tr>
<td>Caltech256</td>
<td>PGM</td>
<td>63.24</td>
<td>1.85</td>
<td></td>
</tr>
</tbody>
</table>