**Semantic-Guided Multi-Attention Localization for Zero-Shot Learning**

Yizhe Zhu1, Jianwen Xie2, Zhiqiang Tang1, Xi Peng3, Ahmed Elgammal1
1 Rutgers University, 2 Hikivision Research Institute, 3 University of Delaware

**Motivation**
Existing zero-shot learning approaches predominantly focus on learning the proper mapping function for visual-semantic embedding, while neglecting the effect of learning discriminative visual features. We observe that multiple discriminative part areas are key points to recognize objects, especially fine-grained objects. For instance, the head and tail are crucial to distinguish bird species.

**Contribution**
- We present a weakly-supervised multi-attention localization model for zero-shot recognition, which jointly discovers the crucial regions and learns feature representation under the guidance of semantic descriptions.
- We propose a multi-attention loss to encourage compact and diverse attention distribution by applying geometric constraints over attention maps.
- We jointly learn global and local features under the supervision of embedding softmax loss and class-center triplet loss to provide an enhanced visual representation for ZSL.
- We conduct extensive experiments and analysis on three zero-shot learning datasets and demonstrate the excellent performance of our proposed method on both part detection and zero-shot learning.

**Zero-shot learning Problem**
Assume there are $N$ labeled instances from $C^o$ seen classes and $D^s$ = \{$(x_i^s, y_i^s, z_i^s)\}_{i=1}^{N^s}$ as training data, where $x_i^s \in X^o$ denotes the image, $y_i^s \in Y^o$ is the corresponding class label, $z_i^s = \phi(y_i^s)$ is $S$ in the semantic representation of the corresponding class. Given an image $x_i^s$ from an unseen class and a set of semantic representations of unseen classes \{z\}, $C^u$ denotes the number of unseen classes, the task of zero-shot learning is to predict the class label $y_i^u$ in $Y^u$ of the image, whose $Y^s$ and $Y^u$ are disjoint.

**Part Detection Results**

<table>
<thead>
<tr>
<th>Method</th>
<th>Head</th>
<th>Tail</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPDA-CNN</td>
<td>90.9</td>
<td>67.2</td>
<td>79.1</td>
</tr>
<tr>
<td>Ours</td>
<td>74.9</td>
<td>48.1</td>
<td>61.5</td>
</tr>
<tr>
<td>Ours w/o MA</td>
<td>65.7</td>
<td>29.4</td>
<td>47.6</td>
</tr>
<tr>
<td>Random</td>
<td>25.6</td>
<td>26.0</td>
<td>25.8</td>
</tr>
</tbody>
</table>

**Method**

- **Multi-Attention Loss**
  To discover compact and diverse regions over attention maps, we design $L_{MA}$:
  \[
  L_{MA} = \sum_{i} \lambda L_{CPT}(M_i) + \lambda L_{DIV}(M_i),
  \]
  where $N_s$ is the number of attention maps, $M_i$ is the $i^{th}$ attention map.
  \[
  L_{CPT}(M_i) = \frac{1}{|M_i - \bar{M_i}|_2^2}
  \]
  where $\bar{M_i}$ is an ideal concentrated attention map, created as a Gaussian blob centering on the peak activation of $M_i$.
  \[
  L_{DIV}(M_i) = \sum_{x \in x_i} m_i^2 \max(0, \hat{m}_i - mrg),
  \]
  where $\hat{m}_i = \max_{x, y \in M_i} m_i^2$ represents the maximum of other attention maps at location $x$ and $mrg$ denotes a margin.

- **Zero-Shot Learning Loss**
  We employ two cooperative losses: the embedding softmax loss to encourage a higher inter-class distinction, and the class-center triplet loss to force the learned feature of each class to be concentrated with a lower intra-class difference.
  \[
  \begin{align*}
  L_{CLS} &= -\frac{1}{N} \log \frac{\exp(s_i)}{\sum_{y_i} \exp(s_j)}, \\
  L_{CCT} &= \max(0, mrg + ||\hat{v}_i - \bar{C}_i||_2 - ||\hat{v}_i - \bar{C}_i||_2) + \lambda C_{T}\end{align*}
  \]
  where $N$ is the number of training samples.

**Zero-Shot Recognition Results**

<table>
<thead>
<tr>
<th>Method</th>
<th>CUB</th>
<th>AWA</th>
<th>FLO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>90.9</td>
<td>61.9</td>
<td>57.7</td>
</tr>
<tr>
<td>Baseline</td>
<td>60.2</td>
<td>41.9</td>
<td>59.8</td>
</tr>
<tr>
<td>Parts</td>
<td>55.4</td>
<td>41.2</td>
<td>59.8</td>
</tr>
<tr>
<td>Baseline+Parts</td>
<td>67.4</td>
<td>64.3</td>
<td>63.9</td>
</tr>
<tr>
<td>Embedding Softmax</td>
<td>60.9</td>
<td>62.4</td>
<td>60.2</td>
</tr>
<tr>
<td>Class-Center Triplet</td>
<td>62.1</td>
<td>64.6</td>
<td>61.1</td>
</tr>
<tr>
<td>Combined</td>
<td>63.5</td>
<td>65.7</td>
<td>61.8</td>
</tr>
</tbody>
</table>

**Reference**