A Shared Multi-Attention Framework for Multi-Label Zero-Shot Learning

Dat Huynh and Ehsan Elhamifar

Khoury College of Computer Sciences
Northeastern University
Motivation

- **Multi-label Learning:**
  - Recognize all labels in an image
  - Require large costly annotations

- **Multi-label Zero-Shot Learning:**
  - Recognize both seen and unseen labels
  - Annotations for only seen labels

- Few work on multi-label ZSL
  - Holistic feature cannot encode all labels
  - Ignore labels from small regions

- **Contributions:**
  - Shared multi-attention features for ZSL
  - Transfers knowledge between seen/unseen
Proposed Architecture

- **Multiple Soft Attention NNs:**
  - Generating *multiple attention feature* for an image

- **Attention Selection** (label-agnostic):
  - For each label, choose attention feature maximizing prediction score

- **Learning:**
  - **Diversity Loss**: Minimize overlap between attention
  - **Relevance Loss**: Focus only on regions improving prediction
  - **Distribution Loss**: Effectively use all attention modules
Experiments

• Recognition: outperforms SOTA on NUS-WIDE and Open Images

<table>
<thead>
<tr>
<th>Method</th>
<th>Task</th>
<th>NUS-WIDE (#seen / #unseen = 925 / 81)</th>
<th>Open Images (#seen / #unseen = 7186 / 400)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$K = 3$</td>
<td>$K = 5$</td>
<td>$K = 10$</td>
</tr>
<tr>
<td></td>
<td>P   R   F1</td>
<td>P   R   F1</td>
<td>P   R   F1</td>
</tr>
<tr>
<td>CONSE</td>
<td>17.5 28.0 21.6</td>
<td>13.9 37.0 20.2</td>
<td>9.4 0.2 7.3</td>
</tr>
<tr>
<td>LabelEM</td>
<td>15.6 25.0 19.2</td>
<td>13.4 35.7 19.5</td>
<td>7.1 0.2 8.7</td>
</tr>
<tr>
<td>Fast0Tag</td>
<td>22.6 36.2 27.8</td>
<td>18.2 48.4 26.4</td>
<td>15.1 0.3 12.6</td>
</tr>
<tr>
<td>One Attention per Label</td>
<td>20.9 33.2 25.8</td>
<td>16.2 43.2 23.6</td>
<td>10.4 - -</td>
</tr>
<tr>
<td>Ours</td>
<td>25.7 41.1 31.6</td>
<td>19.7 52.5 28.7</td>
<td>19.4 0.7 25.6</td>
</tr>
</tbody>
</table>

+3.8% (F1@3) +4.3% (mAP) +0.7% (F1@10) +0.5% (mAP)

• Qualitative Results:

Attention utility depends on label complexity

Successfully attend relevant image regions