CEA LIST at ImageCLEF Scalable Image Concept Annotation 2013

Hervé Le Borgne, Adrian Popescu
GENERAL PRINCIPLE

• Successive « improvements » of the baseline provided
• We define different
  • Visual models $V$ + distance, learning models…
  • Textual models $M$ (tag-based models)
• Principle: for an image to annotate
  • Find visual neighbors of the image into the learning database
  • These (neighboring) images have tags (with a confidence score)
  • The set of tags generates a textual description according to $M$
  • Concepts are described according to $M$ as well $\rightarrow$ similarity
• Late fusion with visual models
• Decision value (0/1) independant for each query
  • Score at 1 when more than average + standard deviation
• Use the development set to test the efficiency of different strategies
FINDING VISUAL NEIGHBORS

• Baseline uses C-SIFT based descriptors
• Our alternative:
  • SIFT, gray-based, densely extracted every 3 pixels
  • Bag of visterm: local soft coding + max pooling
  • Two pyramids:
    • BoV₁ : 1x1 + 3x1 + 2x2
    • BoV₂ : 1x1 + 2x2 + 4x4
• Two distances tested
  • Histogram intersection
  • L₁
• Similar results to the baseline (non significant improvement)
  • K=32 visual neighbors
• We’ll use BoV₁ + L₁
WIKIPEDIA-ESA MODEL

• Explicit Semantic Analysis performed on top of Wikipedia content
  • Map words onto Wikipedia concepts
  • 1187980 wikipedia concepts to start with
  • Concept selection using the inlink count to keep the most frequent concepts – experiments with top 5k concepts
• In the task, map training image annotations to Wikipedia concepts
Inspired by Explicit Semantic Analysis but done with Flickr data

- Flickr 95 – map each image's annotation onto the set of 95 development concepts
- FlickR30k – map each image's annotation onto a set of 30k Wikipedia concepts

Results:

<table>
<thead>
<tr>
<th>Tag</th>
<th>K_{visual}</th>
<th>8</th>
<th>16</th>
<th>32</th>
<th>64</th>
<th>128</th>
</tr>
</thead>
<tbody>
<tr>
<td>co-occurrence</td>
<td>csift</td>
<td>24.71(*)</td>
<td>24.77</td>
<td>24.24</td>
<td>23.63</td>
<td>23.10</td>
</tr>
<tr>
<td>co-occurrence</td>
<td>BoV_{1} + L1</td>
<td>25.01(*)</td>
<td>25.08</td>
<td>24.31</td>
<td>23.60</td>
<td>22.80</td>
</tr>
<tr>
<td>Flickr_{95}</td>
<td>csift</td>
<td>25.08</td>
<td>25.92</td>
<td>26.61</td>
<td>27.33</td>
<td>27.51</td>
</tr>
<tr>
<td>Flickr_{95}</td>
<td>BoV_{1} + L1</td>
<td>25.96</td>
<td>27.30</td>
<td>28.16</td>
<td>28.18</td>
<td>27.67</td>
</tr>
<tr>
<td>FlickR_{30k}</td>
<td>csift</td>
<td>30.25</td>
<td>29.46</td>
<td>29.23</td>
<td>28.80</td>
<td>28.44</td>
</tr>
<tr>
<td>FlickR_{30k}</td>
<td>BoV_{1} + L1</td>
<td>31.05</td>
<td>30.25</td>
<td>29.50</td>
<td>29.07</td>
<td>28.48</td>
</tr>
</tbody>
</table>

Conclusion:

- Significant improvement due to the FlickR_{30k}-based tag model
- The number of visual neighbors considered influences the results in conjunction with the complexity of the tag-based model
LEARNING VISUAL MODELS

• Consider the learning database described with FlickR\textsubscript{30k}

• Principle:
  • Images ranked % score for each concept
  • Select positive and negative samples
  • Learn a SVM-based model for each concept

• Different strategies to choose positive and negative samples
  • S1: 100 most similar Versus 500 least similar
  • S2: two thresholds: positive>0.8 and negative < 0.1
  • S3: usage of visual coherency [Myoupo et al, 2010]

• Results: the simpler, the better!
  • S1: mAP = 0.219
  • S2: mAP = 0.212
  • S2: mAP = 0.209

• Late fusion with tag-based results
PARTICIPATION TO IMAGE CLEF 2013

Run 1: $0.8 \times \text{FlickR}_{30k} + 0.2 \times \text{visual}$
Run 2: $0.8 \times (\text{FlickR}_{30k} + \text{ESA}_{5k}) + 0.2 \times \text{visual}$
Run 3: $0.8 \times (\text{FlickR}_{50k} + \text{ESA}_{5k}) + 0.2 \times \text{visual}$
Run 4: $0.8 \times (\text{FlickR}_{30k} + \text{ESA}_{5k}) + 0.2 \times \text{VC}_{\text{score}}$
Run 4: $0.8 \times (\text{FlickR}_{200k} + \text{ESA}_{5k}) + 0.2 \times \text{VC}_{\text{score}}$

<table>
<thead>
<tr>
<th></th>
<th>devel set</th>
<th>test set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mAP</td>
<td>MF-sample</td>
</tr>
<tr>
<td>Run 1</td>
<td>34.6</td>
<td>28.7</td>
</tr>
<tr>
<td>Run 2</td>
<td>39.6</td>
<td>30.2</td>
</tr>
<tr>
<td>Run 3</td>
<td>40.4</td>
<td>31.8</td>
</tr>
<tr>
<td>Run 4</td>
<td>40.3</td>
<td>32.2</td>
</tr>
<tr>
<td>Run 5</td>
<td>39.2</td>
<td>31.6</td>
</tr>
</tbody>
</table>