The vSTS dataset

Annotation
- Caption pairs annotated for the STS benchmark - Image Description subset.
- Annotators only had access to text.
- Filter out captions referring to the same image (avoid trivial task).

Subsets
- Subset 2014: Subset of the PASCAL VOC-2008 dataset.
  - Obtained 374 pairs (out of 750 in the original file).
- Subset 2015: Subset of Flickr8K benchmark collection for sentence based image description.
  - Obtained 445 pairs (out of 750 in the original file).

Stats

<table>
<thead>
<tr>
<th>Stat</th>
<th>2014</th>
<th>2015</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td># pairs</td>
<td>374</td>
<td>445</td>
<td>819</td>
</tr>
<tr>
<td>mean sim</td>
<td>1.77</td>
<td>1.69</td>
<td>1.72</td>
</tr>
<tr>
<td>std sim</td>
<td>1.49</td>
<td>1.44</td>
<td>1.46</td>
</tr>
<tr>
<td># zeroes</td>
<td>78</td>
<td>81</td>
<td>159</td>
</tr>
</tbody>
</table>

Experiments

Settings
- Dev/Test: Sample 50% at random preserving the overall similarity distribution.
- Train: Part of the text-only STS benchmark dataset as a training set, discarding the examples that overlap with vSTS.
- Evaluation metric: Pearson correlation.

Models
- OVERLAP: Bag-of-words model with cosine similarity.
- CAVERAGE: Glove word embedding based centroid with cosine similarity.
- DAM: Decompositional Attention Model.
- RESNET50: Top layer of a pretrained resnet50 model with cosine similarity.

Combinations
- Combine the predictions of text based models with image based model.
- ⊗: Sum of two outputs.
- ⊕: Multiplication of the output
- LR: Linear regression of two outputs.
  - Parameters estimated with 10fold xval on dev.

Results

<table>
<thead>
<tr>
<th>Modality</th>
<th>Model</th>
<th>Dev set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEXT</td>
<td>A - OVERLAP</td>
<td>0.68</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>B - CAVERAGE</td>
<td>0.65</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>C - DAM</td>
<td>0.71</td>
<td>0.69</td>
</tr>
<tr>
<td>IMAGE</td>
<td>D - RESNET50</td>
<td>0.63</td>
<td>0.61</td>
</tr>
<tr>
<td>Combination</td>
<td>LR</td>
<td>0.77 0.77</td>
<td>0.76 0.75 0.75</td>
</tr>
<tr>
<td>TEXT+IMAGE</td>
<td>A+D</td>
<td>0.77 0.77</td>
<td>0.76 0.75 0.75</td>
</tr>
<tr>
<td></td>
<td>B+D</td>
<td>0.75 0.73</td>
<td>0.76 0.73 0.70</td>
</tr>
<tr>
<td></td>
<td>C+D</td>
<td>0.78 0.78</td>
<td>0.77 0.77 0.78</td>
</tr>
</tbody>
</table>

Discussion

Single models
- DAM obtains the highest Pearson correlation (expected)
- H1 confirmed: Images alone are valid to predict similarity (0.61)

Complementarity
- H2 confirmed: Combination of image and sentence representations obtained the best results (DAM + RESNET50)
- Indications that representation of the real world helps to better understand the text and do better inferences.

Conclusions & Future Work

Contributions
- Creation of dataset of caption pairs with human similarity annotations with access to actual images.
- Test the contribution of visual information in STS.
- Experiments confirmed initial hypotheses.

On going work
- We re-annotated the dataset with scores which are based on both the text and the image.
- First analysis indicate that:
  - Overall similarity values increase when images are present.
  - Similar disagreement on annotators on both settings.
  - High correlation on two annotation frameworks.

Available at
http://ixa2.si.ehu.es/~jiboleo/visual_sts.tgz