Unsupervised Discovery of Multimodal Links in Multi-Image, Multi-Sentence Documents
Jack Hessel, Lillian Lee, David Mimno
Cornell University

What is a "multi-image, multi-sentence document"?

**Image captioning/tagging case**
- single image, explicit multimodal link by construction

**Our case**
- Multiple images, multiple sentences, no explicit multimodal links

Web documents look less like this and more like this!

Multi-image, multi-sentence document use-cases:
1) provide context-specific image captions for low-vision and blind users
2) train image/text models directly from unstructured web documents

The Task: Unsupervised Link Prediction

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Great day at the park</td>
<td>Palyed frisbee with the dog</td>
</tr>
<tr>
<td>Wet out ultimate frisbee</td>
<td></td>
</tr>
</tbody>
</table>

Datasets

**Crowd-labeled Datasets:**
- Designed to address basic questions about this task

Q: Is this task even possible?
A: Microsoft COCO [1] "Documents"

Q: What if images/ sentences are similar within a document?
A: Descriptions-in-isolation [2]

Q: What if sentences are cohesive?
A: Stories-in-Sequence [2]

Q: What if many sentences do not refer to any image?
A: DII-Stress, a version of DII with 45 distractor sentences

**Web-scraped Datasets:**
- Harder, more realistic cases

RecipeQA [3]
- 9K documents, 88K images
- 6 sentences/8 images per doc

"Do it Yourself"[4]
- 9K documents, 154K images
- 15 sentences/16 images per doc

RecipeQA [3]
- 16K documents, 92K images
- 86 sentences/8 images per doc

What makes a document easier or harder?

**Spread Hypothesis:**
- Documents with similar sentences/images will be harder to predict at test-time.

**Content Hypothesis:**
- Some concepts are harder for image/text models to learn.

VS: cat
VS: beautiful

For crowd-labeled datasets, both the spread and content hypothesis explain document difficulty!

Data and Code Available!

http://www.cs.cornell.edu/~jhessel/multiretrieval/multiretrieval.html

---