Multimodal Grounding from User-generated Web Content

Jack Hessel
PhD Candidate
Cornell University
Why study multimodal web data?
Why study multimodal web data?

Increasingly, our social interactions manifest in online, and online communities are increasingly multimodal.
Proportion of Submissions

- Media
- Reddit Text
- Reddit Title
- News


Sources: cnn.com, bbc.co.uk, imgur.com, youtube...
Perhaps the meteoric rise of multimodal content isn't so surprising...
Perhaps the meteoric rise of multimodal content isn't so surprising...

Semioticians have long argued multimodality is a fundamental part of communication

[Lemke 2002]

"The power of visual communication is multiplied when it is co-deployed with language in multimodal texts."
Perhaps the meteoric rise of multimodal content isn't so surprising...

**Semioticians** have long argued multimodality is a fundamental part of communication

[Lemke 2002]

"The power of visual communication is multiplied when it is co-deployed with language in multimodal texts."

**Cognitive psychologists** have studied the connection between perceptions and language since at least the 1970s.

[Miller and Johnson-Laird 1976]
Why study multimodal web data?

Increasingly, our social interactions manifest in online, and online communities are increasingly multimodal
Why study multimodal web data?

Increasingly, our social interactions manifest in *online*, and online communities are increasingly multimodal.

Study of web data gives an in-vivo perspective on communication and communities!

[Vempala and Preotiuc-Pietro 2019; c.f. Chen et al., 2015, Kruk and Lubin et al., 2019, Alikhani et al., 2019]
Why study multimodal web data?
Why study multimodal web data?

if you don't care about communication in online communities
Why study multimodal web data?

if you don't care about communication in online communities

Unreasonable effectiveness of (multimodal) web data

[Deng et al. 2009; Wang et al. 2019; Zhukov et al. 2019]
Why study multimodal web data?

if you don't care about communication in online communities

Unreasonable effectiveness of (multimodal) web data

Building tools that require grounding

[Deng et al. 2009; Wang et al. 2019; Zhukov et al. 2019]

[c.f. Wu et al. 2017; Sharma et al. 2019]
Today
Today

Does *multimodality* affect community reception of content?

[WWW 2017, H., Lee, Mimno]
Today

Does *multimodality* affect community reception of content?

[WWW 2017, H., Lee, Mimno]

What concepts are "groundable," and in what context?

[NAACL 2018, H., Mimno, Lee]
Does multimodality affect community reception of content?

[WWW 2017, H., Lee, Mimno]

What concepts are "groundable," and in what context?

[NAACL 2018, H., Mimno, Lee]

Can grounding be learned directly from multi-sentence, multi-image web documents?

[EMNLP 2019, H., Lee, Mimno]
Today

Does multimodality affect community reception of content?

[WWW 2017, H., Lee, Mimno]

What concepts are "groundable," and in what context?

[NAACL 2018, H., Mimno, Lee]

Can grounding be learned directly from multi-sentence, multi-image web documents?

[EMNLP 2019, H., Lee, Mimno]
Goal:
Recover Community Preferences
by predicting popularity
of image/text posts

"Tonight, I carved a pumpkin. I also doused it in lighter fluid and lit it on fire." - /r/pics

"Snacks!" - /r/aww

"You have to go to the border for food Fish Tacos [San Diego]" - /r/FoodPorn

"Glamor Leaves" - /r/RedditLaquersitas

<table>
<thead>
<tr>
<th></th>
<th># Users</th>
<th>#/% Imgur</th>
<th>Cap Len</th>
</tr>
</thead>
<tbody>
<tr>
<td>pics</td>
<td>2108K</td>
<td>2472K/70%</td>
<td>9.84</td>
</tr>
<tr>
<td>aww</td>
<td>1010K</td>
<td>954K/81%</td>
<td>9.13</td>
</tr>
<tr>
<td>cats</td>
<td>109K</td>
<td>100K/73%</td>
<td>8.97</td>
</tr>
<tr>
<td>MakeupAddiction (MA)</td>
<td>77K</td>
<td>58K/57%</td>
<td>13.67</td>
</tr>
<tr>
<td>FoodPorn (FP)</td>
<td>74K</td>
<td>50K/77%</td>
<td>9.39</td>
</tr>
<tr>
<td>RedditLaqueristas (RL)</td>
<td>27K</td>
<td>39K/73%</td>
<td>11.12</td>
</tr>
</tbody>
</table>
Complication: Minutes matter
Complication: Identity Matters
Complication: Identity Matters

![Graph showing mean score over submission number for MakeupAddiction, cats, RedditLaqueristas, and FoodPorn categories.](image-url)
Complication: Rich get richer

[Salganik et al. 2006]
Complication: Rich get richer

"... small, random rating manipulations on social media submissions created significant changes in downstream ratings... Positive treatment resulted in [an] increased final rating [of] 11.02% on average."
-- Glenski et al. 2015

Reddit overlooked 52% of the most popular links the first time they were submitted.
-- Gilbert 2013
Can we isolate the effects of content rather than context?
Idea: impose strict timing controls
Idea: impose strict timing controls

The grass is always greener

This is why you get two cats
Idea: impose strict timing controls

The grass is always greener
This is why you get two cats

13 Seconds Apart!
Idea: impose strict timing controls

The grass is always greener

This is why you get two cats

13 Seconds Apart!
The grass is always greener  

This is why you get two cats

<table>
<thead>
<tr>
<th></th>
<th>aww</th>
<th>pics</th>
<th>cats</th>
<th>MA</th>
<th>FP</th>
<th>RL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Humans</td>
<td>60.0</td>
<td>63.6</td>
<td>59.6</td>
<td>62.2</td>
<td>72.7</td>
<td>67.2</td>
</tr>
<tr>
<td></td>
<td>Max/Avg Win</td>
<td>Med/Avg Diff</td>
<td># Pairs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>-------------</td>
<td>--------------</td>
<td>---------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pics</td>
<td>30/15 sec</td>
<td>117/478</td>
<td>44K</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>aww</td>
<td>30/15 sec</td>
<td>90/393</td>
<td>33K</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cats</td>
<td>15/7 min</td>
<td>69/231</td>
<td>15K</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MA</td>
<td>60/24 min</td>
<td>88/227</td>
<td>10K</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FP</td>
<td>120/53 min</td>
<td>62/188</td>
<td>8K</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RL</td>
<td>30/14 min</td>
<td>56/118</td>
<td>9K</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Timing</td>
<td>aww</td>
<td>pics</td>
<td>cats</td>
<td>MA</td>
<td>FP</td>
<td>RL</td>
</tr>
<tr>
<td>--------</td>
<td>-----</td>
<td>------</td>
<td>------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Random</td>
<td>50.0</td>
<td>50.0</td>
<td>50.0</td>
<td>50.0</td>
<td>50.0</td>
<td>50.0</td>
</tr>
<tr>
<td>Earlier</td>
<td>51.7</td>
<td>51.1</td>
<td>49.9</td>
<td>48.9</td>
<td>48.6</td>
<td>48.7</td>
</tr>
<tr>
<td>Time</td>
<td>50.2</td>
<td>50.2</td>
<td>50.7</td>
<td>50.4</td>
<td>49.7</td>
<td>50.6</td>
</tr>
</tbody>
</table>
Machine learning experiments

The grass is always greener

This is why you get two cats

$f$

$+1/-1$
Unimodal Results (crossval)
### Unimodal Results (crossval)

<table>
<thead>
<tr>
<th>Type</th>
<th>awm</th>
<th>pics</th>
<th>cats</th>
<th>MA</th>
<th>FP</th>
<th>RL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activity</td>
<td>50.6</td>
<td>51.2</td>
<td>50.7</td>
<td>52.8</td>
<td>51.8</td>
<td>56.1</td>
</tr>
<tr>
<td>Quality</td>
<td>51.1</td>
<td>53.6</td>
<td>52.8</td>
<td>55.0</td>
<td>53.9</td>
<td>60.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type</th>
<th>awm</th>
<th>pics</th>
<th>cats</th>
<th>MA</th>
<th>FP</th>
<th>RL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Struct</td>
<td>54.7</td>
<td>55.5</td>
<td>52.9</td>
<td>60.7</td>
<td>55.5</td>
<td>67.3</td>
</tr>
<tr>
<td>Topic</td>
<td>55.2</td>
<td>55.8</td>
<td>56.8</td>
<td>60.4</td>
<td>55.2</td>
<td>55.5</td>
</tr>
<tr>
<td>DAN</td>
<td>58.6</td>
<td>58.3</td>
<td>58.5</td>
<td>62.2</td>
<td>57.6</td>
<td>59.8</td>
</tr>
<tr>
<td>LSTM</td>
<td>59.4</td>
<td>58.8</td>
<td>58.7</td>
<td>61.0</td>
<td>57.0</td>
<td>59.1</td>
</tr>
<tr>
<td>Bi-LSTM</td>
<td>59.7</td>
<td>58.9</td>
<td>59.3</td>
<td>61.8</td>
<td>57.8</td>
<td>59.6</td>
</tr>
<tr>
<td>Unigram</td>
<td>59.7</td>
<td>58.6</td>
<td>59.5</td>
<td>63.0</td>
<td>57.6</td>
<td>60.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type</th>
<th>awm</th>
<th>pics</th>
<th>cats</th>
<th>MA</th>
<th>FP</th>
<th>RL</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOG</td>
<td>51.7</td>
<td>52.8</td>
<td>51.9</td>
<td>53.5</td>
<td>53.5</td>
<td>53.5</td>
</tr>
<tr>
<td>GIST</td>
<td>52.7</td>
<td>53.0</td>
<td>53.5</td>
<td>55.9</td>
<td>56.5</td>
<td>56.3</td>
</tr>
<tr>
<td>ColorHist</td>
<td>55.3</td>
<td>53.7</td>
<td>55.6</td>
<td>55.0</td>
<td>56.5</td>
<td>54.5</td>
</tr>
<tr>
<td>VGG-19</td>
<td>63.4</td>
<td>58.9</td>
<td>61.1</td>
<td>62.4</td>
<td>62.8</td>
<td>62.1</td>
</tr>
<tr>
<td>ResNet50</td>
<td><strong>64.8</strong></td>
<td><strong>60.0</strong></td>
<td><strong>62.6</strong></td>
<td><strong>64.9</strong></td>
<td><strong>65.2</strong></td>
<td><strong>64.2</strong></td>
</tr>
</tbody>
</table>
Unimodal Results (crossval)

<table>
<thead>
<tr>
<th></th>
<th>aww</th>
<th>pics</th>
<th>cats</th>
<th>MA</th>
<th>FP</th>
<th>RL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type</strong></td>
<td>50.6</td>
<td>51.2</td>
<td>50.7</td>
<td>52.8</td>
<td>51.8</td>
<td>56.1</td>
</tr>
<tr>
<td><strong>Activity</strong></td>
<td>51.1</td>
<td>53.6</td>
<td>52.8</td>
<td>55.0</td>
<td>53.9</td>
<td>60.6</td>
</tr>
<tr>
<td><strong>Quality</strong></td>
<td>54.7</td>
<td>55.5</td>
<td>52.9</td>
<td>60.7</td>
<td>55.5</td>
<td>67.3</td>
</tr>
<tr>
<td><strong>Struct</strong></td>
<td>56.2</td>
<td>54.8</td>
<td>56.5</td>
<td>50.9</td>
<td>52.3</td>
<td>52.5</td>
</tr>
<tr>
<td><strong>Topic</strong></td>
<td>55.2</td>
<td>55.8</td>
<td>56.8</td>
<td>60.4</td>
<td>55.2</td>
<td>55.5</td>
</tr>
<tr>
<td><strong>DAN</strong></td>
<td>58.6</td>
<td>58.3</td>
<td>58.5</td>
<td>62.2</td>
<td>57.6</td>
<td>59.8</td>
</tr>
<tr>
<td><strong>LSTM</strong></td>
<td>59.4</td>
<td>58.8</td>
<td>58.7</td>
<td>61.0</td>
<td>57.0</td>
<td>59.1</td>
</tr>
<tr>
<td><strong>Bi-LSTM</strong></td>
<td>59.7</td>
<td>58.9</td>
<td>59.3</td>
<td>61.8</td>
<td>57.8</td>
<td>59.6</td>
</tr>
<tr>
<td><strong>Unigram</strong></td>
<td>59.7</td>
<td>58.6</td>
<td>59.5</td>
<td>63.0</td>
<td>57.6</td>
<td>60.8</td>
</tr>
<tr>
<td><strong>HOG</strong></td>
<td>51.7</td>
<td>52.8</td>
<td>51.9</td>
<td>53.5</td>
<td>53.5</td>
<td>53.5</td>
</tr>
<tr>
<td><strong>GIST</strong></td>
<td>52.7</td>
<td>53.0</td>
<td>53.5</td>
<td>55.9</td>
<td>56.5</td>
<td>56.3</td>
</tr>
<tr>
<td><strong>ColorHist</strong></td>
<td>55.3</td>
<td>53.7</td>
<td>55.6</td>
<td>55.0</td>
<td>56.5</td>
<td>54.5</td>
</tr>
<tr>
<td><strong>VGG-19</strong></td>
<td>63.4</td>
<td>58.9</td>
<td>61.1</td>
<td>62.4</td>
<td>62.8</td>
<td>62.1</td>
</tr>
<tr>
<td><strong>ResNet50</strong></td>
<td>64.8</td>
<td>60.0</td>
<td>62.6</td>
<td>64.9</td>
<td>65.2</td>
<td>64.2</td>
</tr>
</tbody>
</table>

**User Features**

**Text Features**

**Image Features**
Multimodal Results (crossval)
Multimodal Results (crossval)

<table>
<thead>
<tr>
<th></th>
<th>aww</th>
<th>pics</th>
<th>cats</th>
<th>MA</th>
<th>FP</th>
<th>RL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time + User</td>
<td>54.1</td>
<td>54.7</td>
<td>52.1</td>
<td>58.8</td>
<td>54.2</td>
<td>64.8</td>
</tr>
<tr>
<td>All User</td>
<td>56.3</td>
<td>55.3</td>
<td>54.6</td>
<td>60.9</td>
<td>56.0</td>
<td><strong>68.4</strong></td>
</tr>
<tr>
<td>ResNet50</td>
<td>64.8</td>
<td>60.0</td>
<td>62.6</td>
<td>64.9</td>
<td>65.2</td>
<td>64.2</td>
</tr>
<tr>
<td>Text + Image</td>
<td><strong>67.1</strong></td>
<td><strong>62.7</strong></td>
<td><strong>65.9</strong></td>
<td><strong>67.7</strong></td>
<td><strong>65.8</strong></td>
<td>66.4</td>
</tr>
</tbody>
</table>
## Multimodal Results (crossval)

<table>
<thead>
<tr>
<th></th>
<th>aww</th>
<th>pics</th>
<th>cats</th>
<th>MA</th>
<th>FP</th>
<th>RL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time + User</td>
<td>54.1</td>
<td>54.7</td>
<td>52.1</td>
<td>58.8</td>
<td>54.2</td>
<td>64.8</td>
</tr>
<tr>
<td>All User</td>
<td>56.3</td>
<td>55.3</td>
<td>54.6</td>
<td>60.9</td>
<td>56.0</td>
<td><strong>68.4</strong></td>
</tr>
<tr>
<td>ResNet50</td>
<td>64.8</td>
<td>60.0</td>
<td>62.6</td>
<td>64.9</td>
<td>65.2</td>
<td>64.2</td>
</tr>
<tr>
<td>Text + Image</td>
<td><strong>67.1</strong></td>
<td><strong>62.7</strong></td>
<td><strong>65.9</strong></td>
<td><strong>67.7</strong></td>
<td><strong>65.8</strong></td>
<td>66.4</td>
</tr>
</tbody>
</table>

*Best unimodal*
Multimodal Results (crossval)

<table>
<thead>
<tr>
<th></th>
<th>aww</th>
<th>pics</th>
<th>cats</th>
<th>MA</th>
<th>FP</th>
<th>RL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time + User</td>
<td>54.1</td>
<td>54.7</td>
<td>52.1</td>
<td>58.8</td>
<td>54.2</td>
<td>64.8</td>
</tr>
<tr>
<td>All User</td>
<td>56.3</td>
<td>55.3</td>
<td>54.6</td>
<td>60.9</td>
<td>56.0</td>
<td><strong>68.4</strong></td>
</tr>
<tr>
<td>ResNet50</td>
<td>64.8</td>
<td>60.0</td>
<td>62.6</td>
<td>64.9</td>
<td>65.2</td>
<td>64.2</td>
</tr>
<tr>
<td>Text + Image</td>
<td><strong>67.1</strong></td>
<td><strong>62.7</strong></td>
<td><strong>65.9</strong></td>
<td><strong>67.7</strong></td>
<td><strong>65.8</strong></td>
<td>66.4</td>
</tr>
</tbody>
</table>

Best unimodal
Multimodal beats unimodal!
Multimodal Results (crossval + fully heldout)

<table>
<thead>
<tr>
<th></th>
<th>aww</th>
<th>pics</th>
<th>cats</th>
<th>MA</th>
<th>FP</th>
<th>RL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time + User</td>
<td>54.1</td>
<td>54.7</td>
<td>52.1</td>
<td>58.8</td>
<td>54.2</td>
<td>64.8</td>
</tr>
<tr>
<td>All User</td>
<td>56.3</td>
<td>55.3</td>
<td>54.6</td>
<td>60.9</td>
<td>56.0</td>
<td><strong>68.4</strong></td>
</tr>
<tr>
<td>ResNet50</td>
<td>64.8</td>
<td>60.0</td>
<td>62.6</td>
<td>64.9</td>
<td>65.2</td>
<td>64.2</td>
</tr>
<tr>
<td>Text + Image</td>
<td><strong>67.1</strong></td>
<td><strong>62.7</strong></td>
<td><strong>65.9</strong></td>
<td><strong>67.7</strong></td>
<td><strong>65.8</strong></td>
<td>66.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>aww</th>
<th>pics</th>
<th>cats</th>
<th>MA</th>
<th>FP</th>
<th>RL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time + User</td>
<td>55.5</td>
<td>51.7</td>
<td>52.6</td>
<td>56.9</td>
<td>52.8</td>
<td>60.5</td>
</tr>
<tr>
<td>All User</td>
<td>60.4</td>
<td>51.0</td>
<td>54.3</td>
<td><strong>63.1</strong></td>
<td>57.9</td>
<td><strong>66.0</strong></td>
</tr>
<tr>
<td>Text + Image</td>
<td><strong>65.5</strong></td>
<td><strong>66.0</strong></td>
<td><strong>67.3</strong></td>
<td>62.7</td>
<td><strong>62.6</strong></td>
<td>65.4</td>
</tr>
</tbody>
</table>

Best unimodal

Multimodal beats unimodal!
Machine learning experiments

The grass is always greener

This is why you get two cats

\[ f \]

\[ f \]

\[ f \]
Machine learning experiments

This is why you get two cats

Score
Highest Scores

Lowest Scores
Highest Scores

Lowest Scores
Mean Pool

Activation

Dense (1000)

golden_retriever +0.2290 ***
dingo +0.2126 ***
Labrador_retriever +0.1960 ***
worm_fence +0.1864 ***
cheetah +0.1851 ***
Tibetan_mastiff +0.1830 ***
...
Scotch_terrier -0.2193 ***
bassinet -0.2196 ***
wardrobe -0.2231 ***
miniature_schnauzer -0.2343 ***
four-poster -0.2841 ***
mosquito_net -0.2936 ***

(Significant after applying bonferroni correction)
Machine learning experiments

This is why you get two cats
More evidence that controls are important: our models transfer well to other domains!

[Ding et al. 2019's Instagram results]
Does *multimodality* affect community reception of content?  

[WWW 2017, H., Lee, Mimno]

What concepts are "groundable," and in what context?  

[NAACL 2018, H., Mimno, Lee]

Can grounding be learned directly from multi-sentence, multi-image web documents?  

[EMNLP 2019, H., Lee, Mimno]
Does multimodality affect community reception of content?

[WWW 2017, H., Lee, Mimno]

What concepts are "groundable," and in what context?

[NAACL 2018, H., Mimno, Lee]

Can grounding be learned directly from multi-sentence, multi-image web documents?

[EMNLP 2019, H., Lee, Mimno]
"Performance advantages of [multi-modal approaches] over language-only models have been clearly established when models are required to learn **concrete noun concepts.**" [Hill and Korhonen 2014]
Many datasets focus only on literal objects/actions...

"Do not describe what a person might say."
--- MSCOCO caption annotation guideline for mechanical turkers

... but we encounter lots of non-concrete language on the web!
Work on identifying hard/easy-to-ground concepts:

[Lu et al., 2008; Berg et al., 2010; Parikh and Grauman, 2011; Young et al., 2014; Kiela and Bottou, 2014; Jas and Parikh, 2015; Lazaridou et al., 2015; Silberer et al., 2016; Lu et al., 2017; Bhaskar et al., 2017; Mahajan et al., 2018; inter alia]

Our contributions:

- Fast algorithm for computing concreteness
- Extension from unigrams/bigrams to LDA topics
- Demonstration that concreteness is context specific
The **cat** is in the grass.

This **cat** is enjoying the sun.
The **cat** is in the grass.

This **cat** is enjoying the sun.

This is a **beautiful** baby.

The sunset is **beautiful**.
Conv Net

Image Feature Space

Beautiful

Cat
Measure the clusteredness of concepts by computing expected nearest neighbor concept overlap.
Measure the clusteredness of concepts by computing expected nearest neighbor concept overlap.

Neighbors of beautiful are unlikely to also be beautiful.

Neighbors of cat are likely to be cats.
Connection to Geospatial Statistics

Local Indicators of Spatial Association—LISA

The capabilities for visualization, rapid data retrieval, and manipulation in geographic information systems (GIS) have created the need for new techniques of exploratory data analysis that focus on the “spatial” aspects of the data. The identification of local patterns of spatial association is an important concern in this respect. In this paper, I outline a new general class of local indicators of spatial association (LISA) and show how they allow for the decomposition of global indicators, such as Moran’s I, into the contribution of each observation.

[Amelin 1995]
"Clusteredness" ≈ Concreteness
COCO Results

The man at bat readies to swing at the pitch while the umpire looks on.
# COCO Results

## Most concrete

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>wok</td>
<td>315.595</td>
</tr>
<tr>
<td>hummingbird</td>
<td>291.804</td>
</tr>
<tr>
<td>vane</td>
<td>290.037</td>
</tr>
<tr>
<td>racer</td>
<td>269.043</td>
</tr>
<tr>
<td>grizzly</td>
<td>229.274</td>
</tr>
<tr>
<td>equestrian</td>
<td>219.894</td>
</tr>
<tr>
<td>taxiing</td>
<td>205.410</td>
</tr>
<tr>
<td>unripe</td>
<td>201.733</td>
</tr>
<tr>
<td>siamese</td>
<td>199.024</td>
</tr>
<tr>
<td>delta</td>
<td>195.618</td>
</tr>
<tr>
<td>kiteboarding</td>
<td>192.459</td>
</tr>
<tr>
<td>airways</td>
<td>183.971</td>
</tr>
<tr>
<td>compartments</td>
<td>182.015</td>
</tr>
<tr>
<td>burners</td>
<td>180.553</td>
</tr>
<tr>
<td>stocked</td>
<td>177.472</td>
</tr>
<tr>
<td>spire</td>
<td>177.396</td>
</tr>
<tr>
<td>tulips</td>
<td>173.850</td>
</tr>
<tr>
<td>ben</td>
<td>171.936</td>
</tr>
</tbody>
</table>
COCO Results

Most concrete

- wok      315.595
- hummingbird  291.804
- vane     290.037
- racer    269.043
- grizzly  229.274
- equestrian 219.894
- taxiing  205.410
- unripe   201.733
- siamese  199.024
  delta   195.618
- kiteboarding 192.459
- airways  183.971
- compartments 182.015
- burners  180.553
- stocked  177.472
- spire    177.396
- tulips   173.850
- ben      171.936
COCO Results

Most concrete

- wok: 315.595
- hummingbird: 291.804
- vane: 290.037
- racer: 269.043
- grizzly: 229.274
- equestrian: 219.894
- taxiing: 205.410
  - **unripe**: 201.733
- siamese: 199.024
- delta: 195.618
- kiteboarding: 192.459
- airways: 183.971
- compartments: 182.015
- burners: 180.553
- stocked: 177.472
- spire: 177.396
- tulips: 173.850
- ben: 171.936
COCO Results

Most concrete

<table>
<thead>
<tr>
<th>Term</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>wok</td>
<td>315.595</td>
</tr>
<tr>
<td>hummingbird</td>
<td>291.804</td>
</tr>
<tr>
<td>vane</td>
<td>290.037</td>
</tr>
<tr>
<td>racer</td>
<td>269.043</td>
</tr>
<tr>
<td>grizzly</td>
<td>229.274</td>
</tr>
<tr>
<td>equestrian</td>
<td>219.894</td>
</tr>
<tr>
<td>taxiing</td>
<td>205.410</td>
</tr>
<tr>
<td>unripe</td>
<td>201.733</td>
</tr>
<tr>
<td>siamese</td>
<td>199.024</td>
</tr>
<tr>
<td>delta</td>
<td>195.618</td>
</tr>
<tr>
<td>kiteboarding</td>
<td>192.459</td>
</tr>
<tr>
<td>airways</td>
<td>183.971</td>
</tr>
<tr>
<td>compartments</td>
<td>182.015</td>
</tr>
<tr>
<td>burners</td>
<td>180.553</td>
</tr>
<tr>
<td>stocked</td>
<td>177.472</td>
</tr>
<tr>
<td>spire</td>
<td>177.396</td>
</tr>
<tr>
<td>tulips</td>
<td>173.850</td>
</tr>
<tr>
<td>ben</td>
<td>171.936</td>
</tr>
</tbody>
</table>
## COCO Results

<table>
<thead>
<tr>
<th>Most concrete</th>
<th>Somewhat concrete</th>
<th>Not concrete</th>
</tr>
</thead>
<tbody>
<tr>
<td>wok</td>
<td>motorcycle</td>
<td>side</td>
</tr>
<tr>
<td>hummingbird</td>
<td>fun</td>
<td>while</td>
</tr>
<tr>
<td>vane</td>
<td>including</td>
<td>other</td>
</tr>
<tr>
<td>racer</td>
<td>lays</td>
<td>sits</td>
</tr>
<tr>
<td>grizzly</td>
<td>fish</td>
<td>for</td>
</tr>
<tr>
<td>equestrian</td>
<td>goes</td>
<td>behind</td>
</tr>
<tr>
<td>taxiing</td>
<td>blurry</td>
<td>his</td>
</tr>
<tr>
<td>unripe</td>
<td>helmet</td>
<td>as</td>
</tr>
<tr>
<td>siamese</td>
<td>itself</td>
<td>image</td>
</tr>
<tr>
<td>delta</td>
<td>umbrellas</td>
<td>holding</td>
</tr>
<tr>
<td>kiteboarding</td>
<td>teddy</td>
<td>this</td>
</tr>
<tr>
<td>airways</td>
<td>bar</td>
<td>picture</td>
</tr>
<tr>
<td>compartments</td>
<td>fancy</td>
<td>couple</td>
</tr>
<tr>
<td>burners</td>
<td>sticks</td>
<td>from</td>
</tr>
<tr>
<td>stocked</td>
<td>himself</td>
<td>large</td>
</tr>
<tr>
<td>spire</td>
<td>take</td>
<td>person</td>
</tr>
<tr>
<td>tulips</td>
<td>steps</td>
<td>looking</td>
</tr>
<tr>
<td>ben</td>
<td>attempting</td>
<td>out</td>
</tr>
</tbody>
</table>
"Clusteredness" \approx Concreteness
"Clusteredness" ≈ Concreteness
Context matters!

"London"
Top 1% Concrete
as a caption descriptor in MSCOCO.

"#London"
Rank 1110/7K Concreteness
as a hashtag in a Flickr image
tagging dataset.
Experiments on Wikipedia with LDA topics:

Most Concrete

170.2  hockey
148.9  tennis
86.3   nintendo
81.9   guns
80.9   baseball
76.7   wrestling1
71.4   wrestling2
70.4   software
60.9   auto racing
58.8   currency

Least Concrete

australia  1.95
mexico     1.81
police     1.73
law        1.71
male names 1.65
community  1.58
history    1.52
time       1.47
months     1.43
linguistics 1.29
More concrete = easier to learn
More concrete = easier to learn

(more frequent $\neq$ easier to learn)
Use Case from Shi et al. 2019 (ACL Best Paper Nom.)

Idea: unsupervised constituency parsing based on the concreteness of spans in image captions

<table>
<thead>
<tr>
<th>Model</th>
<th>NP</th>
<th>VP</th>
<th>PP</th>
<th>ADJP</th>
<th>Avg. F1</th>
<th>Self F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>47.3</td>
<td>10.5</td>
<td>17.3</td>
<td>33.5</td>
<td>27.1</td>
<td>32.4</td>
</tr>
<tr>
<td>Left</td>
<td>51.4</td>
<td>1.8</td>
<td>0.2</td>
<td>16.0</td>
<td>23.3</td>
<td>N/A</td>
</tr>
<tr>
<td>Right</td>
<td>32.2</td>
<td>23.4</td>
<td>18.7</td>
<td>14.4</td>
<td>22.9</td>
<td>N/A</td>
</tr>
<tr>
<td>VG-NSL (ours)†</td>
<td>79.6</td>
<td>26.2</td>
<td>42.0</td>
<td>22.0</td>
<td>50.4</td>
<td>87.1</td>
</tr>
<tr>
<td>VG-NSL+HI (ours)†</td>
<td>74.6</td>
<td>32.5</td>
<td>66.5</td>
<td>21.7</td>
<td>53.3</td>
<td>90.2</td>
</tr>
<tr>
<td>VG-NSL+HI+FastText (ours)†</td>
<td>78.8</td>
<td>24.4</td>
<td>65.6</td>
<td>22.0</td>
<td>54.4</td>
<td>89.8</td>
</tr>
<tr>
<td>Hessel et al. (2018)+HI†</td>
<td>72.5</td>
<td>34.4</td>
<td>65.8</td>
<td>26.2</td>
<td>52.9</td>
<td>N/A</td>
</tr>
</tbody>
</table>

(many more baselines in their paper)
Does multimodality affect community reception of content?

[WWW 2017, H., Lee, Mimno]

What concepts are "groundable," and in what context?

[NAACL 2018, H., Mimno, Lee]

Can grounding be learned directly from multi-sentence, multi-image web documents?

[EMNLP 2019, H., Lee, Mimno]
Does *multimodality* affect community reception of content?

[WWW 2017, H., Lee, Mimno]

What concepts are "groundable," and in what context?

[NAACL 2018, H., Mimno, Lee]

Can grounding be learned directly from multi-sentence, multi-image web documents?

[EMNLP 2019, H., Lee, Mimno]
Multi-image, Multi-sentence documents?

Image captioning case:
one image, one sentence
explicit link by annotation

Our case:
multiple images, multiple sentences
no explicit links
Why you might care about multi-image/multi-sentence documents

These types of documents are ubiquitous!

Web pages
Product listings
Books, current and historical
Web comments on images
News articles
...

...
The Task: Unsupervised Link Prediction
What's hard about this link prediction task?

- No explicit labels!
- Sentences may have no image
- Images may have no sentence
- Sentences may have multiple images
- Images may have multiple sentences
Evaluating Link Prediction

Metrics:

AUC: a standard link prediction metric

\[ \sum \sum \mathbb{I}[s(i, j) > s(i', j')] \]

Precision-at-K (we use K=1,5):
"If you had to make your K most confident predictions per-document, how accurate would you be?"
Model

I took the kids down to the river on this fine spring day.

The river has always fascinated me. It’s not a huge river, but it has...

[male] had his adorable hat on, and I loved watching him watch the water.

He found a rock he liked, and asked to take it home.

[male] pointed at everything he saw, and I loved his enthusiasm.
Model
Model

Sentences \[ \hat{M}_i \] Images
Model

maximize \sum_{i,j} \hat{M}_{ij} x_{ij}
Model

\[
\text{maximize } \sum_{i,j} \hat{M}_{ij} x_{ij}
\]

\forall i, \sum_j x_{ij} \leq 1; \forall j, \sum_i x_{ij} \leq 1; \forall i, j, x_{ij} \in \{0, 1\}

to each image, no more than one sentence,
to each sentence, no more than one image
Model

\[
\text{maximize } \sum_{i,j} \hat{M}_{ij} x_{ij}
\]

\[
\forall i, \sum_j x_{ij} \leq 1; \forall j, \sum_i x_{ij} \leq 1; \forall i, j, x_{ij} \in \{0, 1\}
\]

to each image, no more than one sentence, to each sentence, no more than one image

\[
\text{sim}(\hat{M}_i) = \sum_{i,j} \hat{M}_{ij} x_{ij}^*
\]

backprop through the solution $x^*$:
The river has always fascinated me. It’s not a huge river, but it has...

I took the kids down to the river on this fine spring day.

He found a rock he liked, and asked to take it home.

[male] had his adorable hat on, and I loved watching him watch the water.

[male] pointed at everything he saw, and I loved his enthusiasm.

I took the kids down to the river on this fine spring day.

The river has always fascinated me. It’s not a huge river, but it has...
Training: Max Margin loss with Negative Sampling

I took the kids down to the river on this fine spring day. The river has always fascinated me. It's not a huge river, but it has... [male] had his adorable hat on, and I loved watching him watch the water. He found a rock he liked, and asked to take it home. [male] pointed at everything he saw, and I loved his enthusiasm.
Training: Max Margin loss with Negative Sampling

I took the kids down to the river on this fine spring day. The river has always fascinated me. It’s not a huge river, but it has... [male] had his adorable hat on, and I loved watching him watch the water. He found a rock he liked, [male] pointed at everything he saw, and I loved his enthusiasm.
Training: Max Margin loss with Negative Sampling

I took the kids down to the river on this fine spring day. The river has always fascinated me. It's not a huge river, but it has...

[male] had his adorable hat on, and I loved watching him watch the water. He found a rock he liked, and asked to take it home. [male] pointed at everything he saw, and I loved his enthusiasm.
Training: Max Margin loss with Negative Sampling

I took the kids down to the river on this fine spring day. The river has always fascinated me. It's not a huge river, but it has... [male] had his adorable hat on, and I loved watching him watch the water. He found a rock he liked, and asked to take it home. [male] pointed at everything he saw, and I loved his enthusiasm.
I took the kids down to the river on this fine spring day. The river has always fascinated me. It's not a huge river, but it has... [male] had his adorable hat on, and I loved watching him watch the water. He found a rock he liked, and asked to take it home. [male] pointed at everything he saw, and I loved his enthusiasm.
I took the kids down to the river on this fine spring day. The river has always fascinated me. It's not a huge river, but it has...

[male] had his adorable hat on, and I loved watching him watch the water. He found a rock he liked, and asked to take it home. [male] pointed at everything he saw, and I loved his enthusiasm.
I took the kids down to the river on this fine spring day. The river has always fascinated me. It's not a huge river, but it has... [male] had his adorable hat on, and I loved watching him watch the water. He found a rock he liked, and asked to take it home. [male] pointed at everything he saw, and I loved his enthusiasm.

\[
\mathcal{L}(S_i, V_i) = \max_{V' \in \mathcal{V}'} h\left(\text{sim}(S_i, V_i), \text{sim}(S_i, V')\right) + \max_{S' \in \mathcal{S}'} h\left(\text{sim}(S_i, V_i), \text{sim}(S', V_i)\right)
\]
I took the kids down to the river on this fine spring day. The river has always fascinated me. It's not a huge river, but it has... [male] had his adorable hat on, and I loved watching him watch the water. He found a rock he liked, and asked to take it home. [male] pointed at everything he saw, and I loved his enthusiasm.

\[
\mathcal{L} (S_i, V_i) = \max_{V' \in V'} h \left( \text{sim}(S_i, V_i), \text{sim}(S_i, V') \right)
\]

\[
+ \max_{S' \in S'} h \left( \text{sim}(S_i, V_i), \text{sim}(S', V_i) \right)
\]
The paper has results on four crowdsourced datasets but we'll focus on the web-scraped data for now...
Datasets Scraped from the Web
Datasets Scraped from the Web

Data scraped from instructables.com;

Via web interface, authors associate multiple images with recipe steps, which gives us a graph for evaluation.

RecipeQA

Ingredients Mint Layer 1. 1 sticks butter 2. 1 cup powdered sugar 3. 1 table spoon milk ...

*** Chocolate Layer #1 Although the chocolate layers are perhaps the simplest... until smooth

*** Finishing First Layer 1. Pour evenly into a pan... ***

Onto the Mint! The Mint mixture can be changed... Second Layer Is Finished! Now comes a bit of a tricky part. ...The possibilities are endless :D *** Repeat Step #2

... and final layer of your beautiful snack. *** Pulling It All Together! 1. Remove the dually layered bar ...

*** Finishing Notes Allow the bar to acclimate...
So my partner and I decided that we want to build our first In-Home rock climbing wall... *** We set aside a budget of $1200 and began a model to estimate... *** Each box represents one square foot of climbing space... *** After cutting a bit more plywood and lining it up... *** I insisted in putting a few cross braces into the angled section... *** I’m going to have fun with this.
Datasets Scraped from the Web

Rivet

A rivet is a permanent mechanical fastener... Solid rivets consist simply of a shaft and head... Steel rivets can be found in static structures such as bridges, cranes, ... They are offered from 1/16-inch (1.6 mm) to 3/8-inch (9.5 mm) in diameter ... The most common machine is the impact riveter and the most common use of semitubular rivets is in lighting, brakes ...

Data scraped from wikipedia

There are no ground-truth links between images and text (so we are limited to qualitative observation).
# Stats for Web Datasets

<table>
<thead>
<tr>
<th></th>
<th>train/val/test</th>
<th>$n_i/m_i$ (median)</th>
<th># imgs (unique)</th>
<th>density</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIY</td>
<td>7K/1K/1K</td>
<td>15/16</td>
<td>154K</td>
<td>8%</td>
</tr>
<tr>
<td>RQA</td>
<td>7K/1K/1K</td>
<td>6/8</td>
<td>88K</td>
<td>17%</td>
</tr>
<tr>
<td>WIKI</td>
<td>14K/1K/1K</td>
<td>86/5</td>
<td>92K</td>
<td>N/A</td>
</tr>
</tbody>
</table>

# sentences/doc  # images/doc
Quantitative Results on RQA/DIY

<table>
<thead>
<tr>
<th></th>
<th>RQA</th>
<th>DIY</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUC</td>
<td>p@1/p@5</td>
</tr>
<tr>
<td>Random</td>
<td>49.4</td>
<td>17.8/16.7</td>
</tr>
<tr>
<td>Obj Detect</td>
<td>58.7</td>
<td>25.1/21.5</td>
</tr>
<tr>
<td>NoStruct</td>
<td>60.5</td>
<td>33.8/27.0</td>
</tr>
</tbody>
</table>
Quantitative Results on RQA/DIY

<table>
<thead>
<tr>
<th></th>
<th>RQA</th>
<th>DIY</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUC</td>
<td>p@1</td>
</tr>
<tr>
<td>Random</td>
<td>49.4</td>
<td>17.8</td>
</tr>
<tr>
<td>Obj Detect</td>
<td>58.7</td>
<td>25.1</td>
</tr>
<tr>
<td>NoStruct</td>
<td>60.5</td>
<td>33.8</td>
</tr>
<tr>
<td>AP</td>
<td><strong>69.3</strong></td>
<td><strong>47.3</strong></td>
</tr>
<tr>
<td></td>
<td><strong>61.8</strong></td>
<td><strong>22.5</strong></td>
</tr>
</tbody>
</table>
Pour the quart of half-and-half into the blender. Weigh out about 120g...

First, fry up a pound of your favorite thin-sliced bacon. For this dish...

While I made a triple batch for competition, this recipe is scaled...

This layer will be your "meat" strip in the center of the bacon...

This one is just syrup and smoke. Combine 1 cup bacon...

Example from RQA
First sighted by Europeans around 1600 on Mauritius, the dodo became extinct less than eighty years later.

The island is well known for its natural beauty.

... a significant migrant population of Bhumihar Brahmins in Mauritius who have made a mark for themselves in different fields.

Mauritian Créole, which is spoken by 90 per cent of the population, is considered to be the native tongue.

For the dodo, the object detection baseline's selected sentence began with:

“(Mauritian Creole people usually known as ‘Creoles’)”
Does *multimodality* affect community reception of content?

[WWW 2017, H., Lee, Mimno]

What concepts are "groundable," and in what context?

[NAACL 2018, H., Mimno, Lee]

Can grounding be learned directly from multi-sentence, multi-image web documents?

[EMNLP 2019, H., Lee, Mimno]
Ongoing work: exploring grounding in web videos

A Case Study on Combining ASR and Visual Features for Generating Instructional Video Captions

[CoNLL 2019 H., Pang, Zhu, Soricut; H., Pang, Zhu are planning ACL submission :)]
Ongoing work: incorporating structure into multi-retrieval models

Training Data: Document-level Co-occurrence

Objective 1: Image/Sentence Link Prediction
- Great day at the park!
- Played frisbee with the dog.
- Won our ultimate game!

Objective 2: Region/Span Link Prediction
- Played frisbee with the dog.
Ongoing work: decoupling additive vs. multiplicative interactions

This is why you get two cats

Additive signals

Multiplicative signals
Thanks to my awesome collaborators!

Lillian Lee
David Mimno
Bo Pang
Radu Soricut
Zhenhai Zhu
And thanks to you for having me!!

The grass is always greener
This is why you get two cats

"... dogs ..."

"... beautiful ..."

Training Time: Document-level Co-occurrence
Testing Time: Image/Sentence Link Prediction

Contact:
jmhessel@gmail.com
@jmhessel on Twitter

Code, data, and papers are all available:
http://www.cs.cornell.edu/~jhessel/