Vision, Language and Commonsense
What does it mean to “understand”?
Red
Is a sheep fluffy?

A tree?
“Long, blue, spiky-edged shadows crept out across the snow-fields, while a rosy glow, at first scarce discernible, gradually deepened and suffused every mountain-top, flushing the glaciers and the harsh crags above them.”

John Muir
Vision ➔ Representation ➔ Language
A woman is taking a

Show and Tell: A Neural Image Caption Generator,
Vinyals et al., CVPR 2015
A man standing next to a fire hydrant in front of a brick building.

From Captions to Visual Concepts and Back,
Fang et al., CVPR 2015
35% - 85% of captions are identical to training captions.
Nearest Neighbor

Test

Train
Nearest Neighbor

A black and white cat sitting in a bathroom sink.

Two zebras and a giraffe in a field.

See mscoco.org for image information
## MS COCO Caption Challenge

<table>
<thead>
<tr>
<th>Method</th>
<th>CIDEr-D</th>
<th>Meteor</th>
<th>ROUGE-L</th>
<th>BLEU-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google[4]</td>
<td>0.943</td>
<td>0.254</td>
<td>0.53</td>
<td>0.309</td>
</tr>
<tr>
<td>MSR Captivator[9]</td>
<td>0.931</td>
<td>0.248</td>
<td>0.526</td>
<td>0.308</td>
</tr>
<tr>
<td>m-RNN[15]</td>
<td>0.917</td>
<td>0.242</td>
<td>0.521</td>
<td>0.299</td>
</tr>
<tr>
<td>MSR[8]</td>
<td>0.912</td>
<td>0.247</td>
<td>0.519</td>
<td>0.291</td>
</tr>
<tr>
<td>Nearest Neighbor[11]</td>
<td>0.886</td>
<td>0.237</td>
<td>0.507</td>
<td>0.280</td>
</tr>
<tr>
<td>m-RNN (Baidu/ UCLA)[16]</td>
<td>0.886</td>
<td>0.238</td>
<td>0.524</td>
<td>0.302</td>
</tr>
<tr>
<td>Berkeley LRCN[2]</td>
<td>0.869</td>
<td>0.242</td>
<td>0.517</td>
<td>0.277</td>
</tr>
<tr>
<td>Human[5]</td>
<td>0.854</td>
<td>0.252</td>
<td>0.484</td>
<td>0.217</td>
</tr>
<tr>
<td>Montreal/Toronto[10]</td>
<td>0.85</td>
<td>0.243</td>
<td>0.513</td>
<td>0.268</td>
</tr>
<tr>
<td>PicSOM[13]</td>
<td>0.833</td>
<td>0.231</td>
<td>0.505</td>
<td>0.281</td>
</tr>
<tr>
<td>MLBL[7]</td>
<td>0.74</td>
<td>0.219</td>
<td>0.499</td>
<td>0.26</td>
</tr>
<tr>
<td>ACVT[1]</td>
<td>0.709</td>
<td>0.213</td>
<td>0.483</td>
<td>0.246</td>
</tr>
<tr>
<td>NeuralTalk[12]</td>
<td>0.674</td>
<td>0.21</td>
<td>0.475</td>
<td>0.224</td>
</tr>
<tr>
<td>Tsinghua Bigeye[14]</td>
<td>0.673</td>
<td>0.207</td>
<td>0.49</td>
<td>0.241</td>
</tr>
<tr>
<td>MIL[6]</td>
<td>0.666</td>
<td>0.214</td>
<td>0.468</td>
<td>0.216</td>
</tr>
<tr>
<td>Brno University[3]</td>
<td>0.517</td>
<td>0.195</td>
<td>0.403</td>
<td>0.134</td>
</tr>
</tbody>
</table>
Why does nearest neighbor work?
A man riding a wave on a surfboard in the water.
vemödalen - n. the frustration of photographing something amazing when thousands of identical photos already exist

Vemödalen: The Fear That Everything Has Already Been Done
https://www.youtube.com/watch?v=8ftDjebw8aA
Going beyond...

From Captions to Visual Concepts and Back, Fang et al., CVPR 2015.

A girl

A girl and boy
A girl and boy knocked
A girl and boy knocked down the tower.
A girl and boy knocked down the tower.
Vision ↔ Representation ↔ Language
Recurrent Neural Networks

Learning a Recurrent Visual Representation for Image Caption Generation
Chen and Zitnick, CVPR 2015.
Mike is holding a baseball bat.
Jenny just threw the baseball.
Mike did not hit the baseball.
The cat is looking at the hot dog.
Lightning is striking the tree.
Apples are growing on the tree.
Limitations
A crazy zebra climbing a giraffe to get a better view.

The limits of vision and language models...
A man is rescued from his truck that is hanging dangerously from a bridge.
Abstract Scenes

Devi Parikh, VT
Xinlei Chen, CMU
Rama Vedantam, VT
David Fouhey, CMU
Stan Antol, VT
Xiao Lin, VT
Lucy Vanderwende, Microsoft Research
Is photorealism necessary?
Jenny just threw the beach ball angrily at Mike while the dog watches them both.
Mike fights off a bear by giving him a hotdog while jenny runs away.
run after

run to
Learning the Visual Interpretation of Sentences, Zitnick, Parikh, and Vanderwende, ICCV 2013.
Visual Question Answering
VQA: Visual Question Answering

Antol et al., ArXiv, 2015.

visualqa.org
VQA: Visual Question Answering

How many slices of pizza are there? Is this a vegetarian pizza?

Does it appear to be rainy? Does this person have 20/20 vision?

Is this person expecting company? What is just under the tree?
VQA: Visual Question Answering

July 2015 (Beta release)
- 120k images (360k questions, 3.6M answers)

September 2015 (Full release)
- 120k COCO train+val images
- 60k “random” images
- 50k abstract scenes

visualqa.org