Grounding distributional semantics in the visual world

Marco Baroni

Center for Mind/Brain Sciences
University of Trento

VL’15
Lisbon, Portugal
In collaboration with:

Angeliki Lazaridou

Nghia The Pham, Marco Marelli, Raquel Fernandez, Grzegorz Chrupała
What is word meaning made of?
The classical view

man: +HUMAN +MALE +ADULT ±MARRIED

bachelor: +HUMAN +MALE +ADULT −MARRIED

Adapted from Boleda and Erk AAAI 2015
Near synonymy
Edmonds and Hirst CL 2002

man: +HUMAN +MALE +ADULT

gentleman, lad, chap, dude, bloke, guy:
+HUMAN +MALE +ADULT ±???

Adapted from Boleda and Erk AAAI 2015
Distributed representations

man  gentleman  bloke  lad

bachelor  gentleman

guy  chap  man  lad  bloke  dude
Add any liquid left from the ficle together with all the other ingredients except the breadcrumbs and cheese.

(Stimulus from Lazaridou et al. in preparation)
Inducing semantic vectors from context

... it was getting late
the tired \textit{gentleman} sat on the sofa
he hoped the guests would leave soon
...

\textit{gentleman}
**Men in distributed semantic space**

<table>
<thead>
<tr>
<th>man</th>
<th>gentleman</th>
<th>lad</th>
<th>bloke</th>
</tr>
</thead>
<tbody>
<tr>
<td>woman</td>
<td>gentlewoman</td>
<td>boy</td>
<td>chap</td>
</tr>
<tr>
<td>gentleman</td>
<td>Hunsden</td>
<td>bloke</td>
<td>guy</td>
</tr>
<tr>
<td>gray-haired</td>
<td>Lestrade</td>
<td>scouser</td>
<td>toser</td>
</tr>
<tr>
<td>boy</td>
<td>Utterson</td>
<td>lass</td>
<td>twat</td>
</tr>
<tr>
<td>person</td>
<td>Scotchman</td>
<td>youngster</td>
<td>fella</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>chap</th>
<th>dude</th>
<th>guy</th>
<th>bachelor</th>
</tr>
</thead>
<tbody>
<tr>
<td>bloke</td>
<td>freakin’</td>
<td>bloke</td>
<td>bachelor’s</td>
</tr>
<tr>
<td>guy</td>
<td>woah</td>
<td>chap</td>
<td>master’s</td>
</tr>
<tr>
<td>lad</td>
<td>dorky</td>
<td>doofus</td>
<td>doctorate</td>
</tr>
<tr>
<td>fella</td>
<td>dumbass</td>
<td>dude</td>
<td>majoring</td>
</tr>
<tr>
<td>man</td>
<td>stoopid</td>
<td>fella</td>
<td>degree</td>
</tr>
</tbody>
</table>

_http://clic.cimec.unitn.it/composes/semantic-vectors.html_
The grounding problem
The psychedelic world of distributional semantic color

- clover is blue
- coffee is green
- crows are white
- flour is black
- fog is green
- gold is purple
- mud is red
- the sky is green
- violins are blue

Bruni et al. ACL 2012

Disjoint induction of multimodal spaces
Feng and Lapata NAACL 2010, Bruni et al. JAIR 2014…

Lucifer Sam, siam cat. Always sitting by your side
Always by your side. That cat's something I can't
explain. Ginger, ginger, Jennifer Gentle you're a witch.
You're the left side He's the right side. Oh, no! That
cat's something I can't explain. Lucifer go to sea. Be a
hip cat, be a ship's cat. Somewhere, anywhere. That
cat's something I can't explain. At night prowling sifting
sand. Hiding around on the ground. He'll be found
when you're around. That cat's something I can't explain.
The multimodal skip-gram model

Input stream

The cute *cat* sat on the mat

The sad *cow* was looking at us

toss me the *rabbit*!

Wild *horses* couldn’t drag me away

Three little *piggies* went to the market

...
The multimodal skip-gram model
Learning when only linguistic contexts are available

three little *piggies* went to the market

Equivalent to Mikolov et al.’s *skip-gram* (“word2vec”) model
The multimodal skip-gram model
Learning from joint linguistic/visual contexts

the cute *cat* sat on the mat

visual feature extraction

linguistic context prediction

visual feature prediction

semantic vector induction
Approximating human similarity judgments
Figure of merit: Spearman’s ρ

<table>
<thead>
<tr>
<th>examples</th>
<th>MEN</th>
<th>Simlex-999</th>
<th>SemSim</th>
<th>VisSim</th>
</tr>
</thead>
<tbody>
<tr>
<td>bakery bread</td>
<td></td>
<td>happy cheerful</td>
<td>jeans sweater</td>
<td>donkey horse</td>
</tr>
<tr>
<td>Bruni et al.</td>
<td>0.78</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hill et al.</td>
<td></td>
<td>0.41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Silberer and Lapata</td>
<td></td>
<td></td>
<td>0.70</td>
<td>0.64</td>
</tr>
<tr>
<td>visual vectors</td>
<td>0.62*</td>
<td>0.54*</td>
<td>0.55*</td>
<td>0.56*</td>
</tr>
<tr>
<td>linguistic vectors</td>
<td>0.70</td>
<td>0.33</td>
<td>0.62</td>
<td>0.48</td>
</tr>
<tr>
<td>multimodal SVD</td>
<td>0.61</td>
<td>0.28</td>
<td>0.65</td>
<td>0.58</td>
</tr>
<tr>
<td>multimodal skip-gram</td>
<td>0.75</td>
<td>0.37</td>
<td><strong>0.72</strong></td>
<td>0.63</td>
</tr>
</tbody>
</table>
## Nearest neighbour examples

<table>
<thead>
<tr>
<th><strong>language only</strong></th>
<th><strong>multimodal</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>donut</strong></td>
<td>pizza, sushi, sandwich</td>
</tr>
<tr>
<td><strong>owl</strong></td>
<td>eagle, woodpecker, falcon</td>
</tr>
<tr>
<td><strong>mural</strong></td>
<td>painting, portrait, sculpture</td>
</tr>
<tr>
<td><strong>tobacco</strong></td>
<td>cigarette, cigar, corn</td>
</tr>
<tr>
<td><strong>depth</strong></td>
<td>sea, underwater, level</td>
</tr>
<tr>
<td><strong>chaos</strong></td>
<td>demon, anarchy, destruction</td>
</tr>
</tbody>
</table>
Out-of-the-box 0-shot image retrieval with MSG

MSG training

tiger

jaguar

leopard

panther

puma

lynx

lion
Out-of-the box 0-shot image retrieval with MSG
Test-time retrieval
Out-of-the box 0-shot image retrieval with MSG

Search space: 5.1K images with unique labels; percentage precision

<table>
<thead>
<tr>
<th>Method</th>
<th>P@1</th>
<th>P@10</th>
<th>P@20</th>
<th>P@50</th>
</tr>
</thead>
<tbody>
<tr>
<td>chance</td>
<td>&lt;0.1</td>
<td>0.2</td>
<td>0.4</td>
<td>1.0</td>
</tr>
<tr>
<td>skip-gram/supervised cross-modal mapping</td>
<td>2.3</td>
<td>11.9</td>
<td>17.9</td>
<td>30.9</td>
</tr>
<tr>
<td><strong>multimodal skip-gram/direct retrieval</strong></td>
<td>2.0</td>
<td>14.1</td>
<td>20.1</td>
<td>33.0</td>
</tr>
</tbody>
</table>
Nearest visual neighbours of abstract words

freedom  theory  wrong

god  together  place

Subjects’ significant preference for true neighbour over confounder:
random level: 0%
unseen abstract: 23%
unseen concrete: 53%
Abstractness correlates with MSG entropy

\( \rho > 0.7 \) on Kiela et al. ACL 2014 data set, no correlation for skip-gram vectors!
Realistic word learning challenges for MSG
Learn from real conversational data (ideally, child-directed speech)

A hat is a head covering. It can be worn for protection against the elements, ceremonial reason, religious reasons, safety, or as a fashion accessory.
Realistic word learning challenges for MSG
Learn from real conversational data (ideally, child-directed speech)

A hat is a head covering. It can be worn for protection against the elements, ceremonial reason, religious reasons, safety, or as a fashion accessory.

peekaboo
peekaboo
peekaboo
ahhah
ahhah
whos this on the hat
i think this is oh thats minniemouse
do you see minniemouse
yes you see minniemouse
Realistic word learning challenges for MSG

Referential uncertainty

*the cute cat sat on the mat*
Realistic word learning challenges for MSG
Learning from minimal exposure ("fast mapping")
moms got a \textit{hat} on, look
The Frank corpus

http://langcog.stanford.edu/materials/nipsmaterials.html

*mot let me have that
%ref: RING
*mot ahhah what's this
%ref: RING HAT
*mot what does mom look like with the hat on
%ref: RING HAT
*mot do i look pretty good with the hat on
%ref: RING HAT
*mot hmm
%ref: RING HAT
*mot hmm
%ref: RING HAT
*mot do i look pretty good
%ref: RING HAT
*mot peekaboo
%ref: RING HAT
The Frank corpus
Our version

let me have that

ahhah whats this

what does mom look like with the hat on

do i look pretty good with the hat on

hmm
Matching words with objects
36 test words, 17 test objects

<table>
<thead>
<tr>
<th>Model</th>
<th>Best F</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSG</td>
<td>.75</td>
</tr>
<tr>
<td>BEAGLE</td>
<td>.55</td>
</tr>
<tr>
<td>PMI</td>
<td>.53</td>
</tr>
<tr>
<td>Bayesian CSL</td>
<td>.54</td>
</tr>
<tr>
<td>(BEAGLE+PMI)</td>
<td>.83</td>
</tr>
</tbody>
</table>

Bayesian CSL: Frank et al. NIPS 2007
### MSG object identification after a single exposure

<table>
<thead>
<tr>
<th>word</th>
<th>gold object</th>
<th>17 objects</th>
<th>5K objects</th>
</tr>
</thead>
<tbody>
<tr>
<td>bunny</td>
<td>bunny</td>
<td>bunny</td>
<td>hare</td>
</tr>
<tr>
<td>cows</td>
<td>cow</td>
<td>cow</td>
<td>heifer</td>
</tr>
<tr>
<td>duck</td>
<td>duck</td>
<td>hand</td>
<td>chronograph</td>
</tr>
<tr>
<td>duckie</td>
<td>duck</td>
<td>hand</td>
<td>chronograph</td>
</tr>
<tr>
<td>kitty</td>
<td>kitty</td>
<td>kitty</td>
<td>kitten</td>
</tr>
<tr>
<td>lambie</td>
<td>lamb</td>
<td>lamb</td>
<td>lamb</td>
</tr>
<tr>
<td>moocows</td>
<td>cow</td>
<td>pig</td>
<td>bison</td>
</tr>
<tr>
<td>rattle</td>
<td>rattle</td>
<td>hand</td>
<td>invader</td>
</tr>
</tbody>
</table>
THANK YOU!