Grounding word representations in the visual world

Marco Baroni

Center for Mind/Brain Sciences
University of Trento

LEAR (Grenoble)
July 2015

In collaboration with:

Angeliki Lazaridou

Nghia The Pham, Marco Marelli,
Raquel Fernandez, Grzegorz Chrupała,
Dat Tien Nguyen, Raffaella Bernardi

What is word meaning made of?

The classical view

man: $+$HUMAN $+$MALE $+$ADULT $\pm$MARRIED

bachelor: $+$HUMAN $+$MALE $+$ADULT $-$MARRIED

Adapted from Boleda and Erk AAAI 2015

Near synonymy

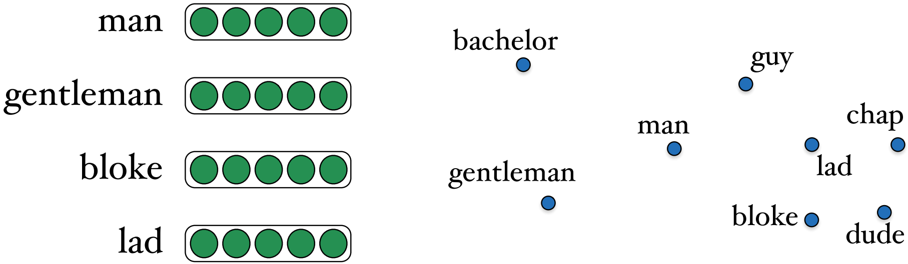
Edmonds and Hirst CL 2002

man: +HUMAN +MALE +ADULT

gentleman, lad, chap, dude, bloke, guy:
+HUMAN +MALE +ADULT $\pm$???

Adapted from Boleda and Erk AAAI 2015

Distributed representations



Context as distant semantic supervision

Distributed and distributional semantics

Add any liquid left from the **ficle** together with all the other ingredients except the breadcrumbs and cheese.

Figure from Lazaridou et al. *in preparation*

Inducing semantic vectors from context

Landauer and Dumais PsychRev 1997, Schütze's 1997 CSLI book, Griffiths et al. PsychRev 2007, Mikolov et al. NIPS 2013

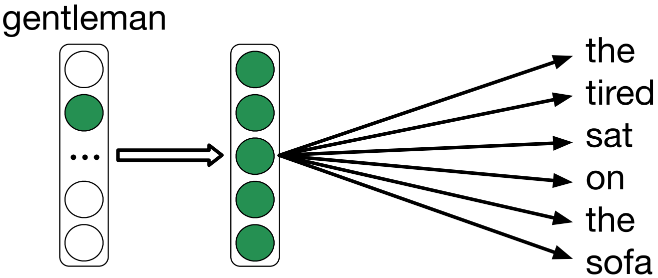
…

his father was a real *gentleman*

the tired *gentleman* sat on the sofa

we met the old *gentleman* in the park

…



Men in distributed semantic space

man	gentleman	lad	bloke
woman	gentlewoman	boy	chap
gentleman	Hunsden	bloke	guy
gray-haired	Lestrade	scouser	tosser
boy	Utterson	lass	twat
person	Scotchman	youngster	fella

chap	dude	guy	bachelor
bloke	freakin’	bloke	bachelor’s
guy	woah	chap	master’s
lad	dorky	doofus	doctorate
fella	dumbass	dude	majoring
man	stoopid	fella	degree

`http://clic.cimec.unitn.it/composes/semantic-vectors.html`

The grounding problem

The psychedelic world of distributional semantic color

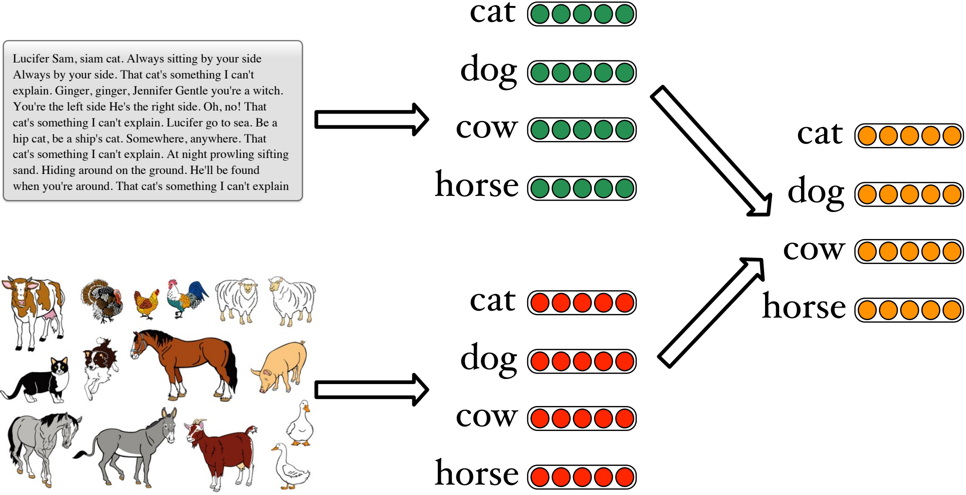
- **clover** is blue
- **coffee** is green
- **crows** are white
- **flour** is **black**
- **fog** is green
- **gold** is purple
- **mud** is red
- the **sky** is green
- **violins** are blue

Bruni et al. ACL 2012

See also: Andrews et al. PsychRev 2009, Baroni et al. CogSciJ 2010, Riordan and Jones TopiCS 2011...

Disjoint induction of multimodal spaces

Feng and Lapata NAACL 2010, Bruni et al. JAIR 2014. . .



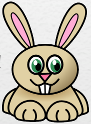
The multimodal skip-gram model

Input stream

the cute *cat* sat on the mat

the sad *cow* was looking at us

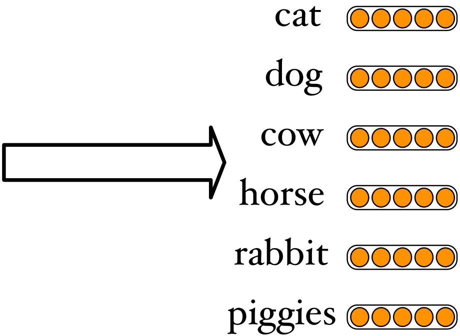
toss me the *rabbit*!



wild *horses* couldn't drag me away

three little *piggies* went to the market

...

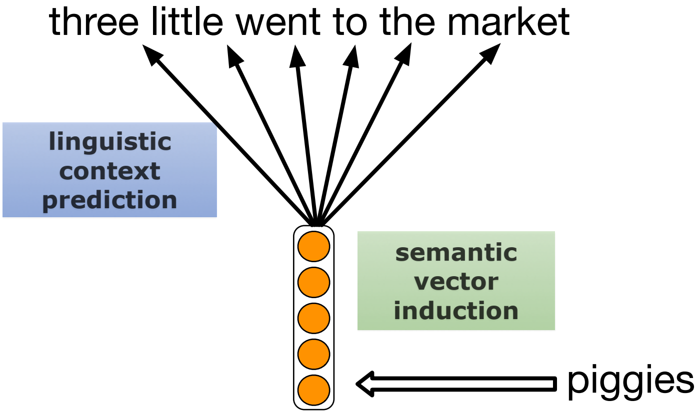


Lazaridou et al. NAACL 2015

The multimodal skip-gram model

Learning when only linguistic contexts are available

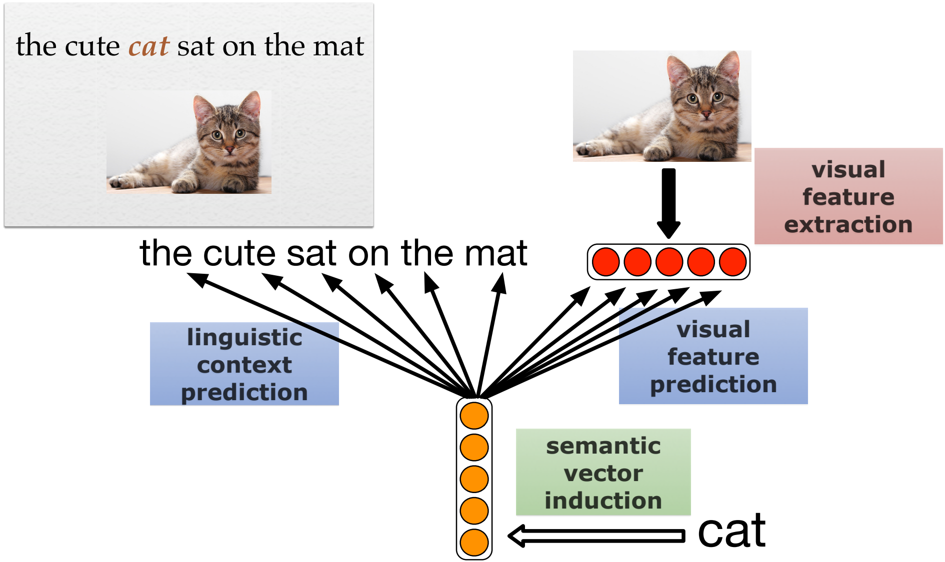
three little *piggies* went to the market



Equivalent to Mikolov et al.'s *skip-gram* ("word2vec") model

The multimodal skip-gram model

Learning from joint linguistic/visual contexts



Approximating human similarity judgments

Figure of merit: Spearman's ρ

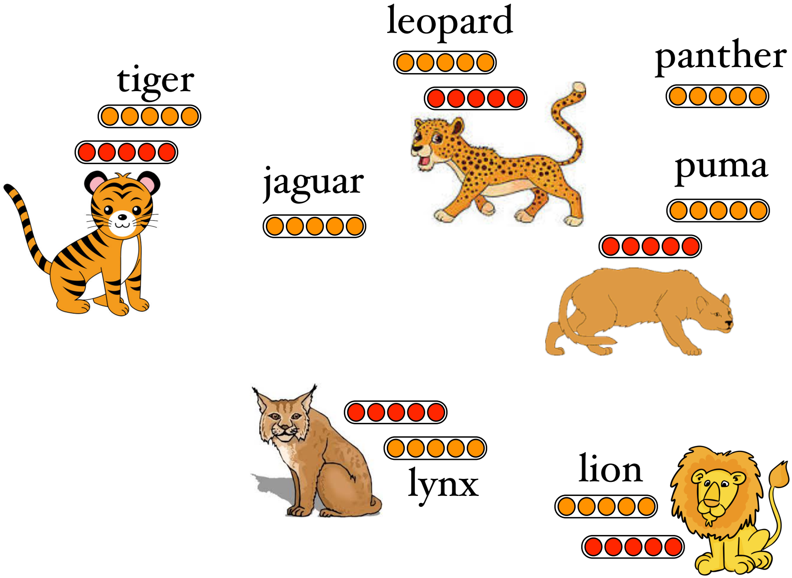
	MEN	Simlex-999	SemSim	VisSim
examples	bakery bread	happy cheerful	jeans sweater	donkey horse
Bruni et al.	**0.78**			
Hill et al.		**0.41**		
Silberer and Lapata			0.70	**0.64**
visual vectors	0.62*	0.54*	0.55*	0.56*
linguistic vectors	0.70	0.33	0.62	0.48
multimodal SVD	0.61	0.28	0.65	0.58
multimodal skip-gram	0.75	0.37	**0.72**	0.63

Nearest neighbour examples

	language only	*multimodal*
donut	fridge, diner, candy	pizza, sushi, sandwich
owl	pheasant, woodpecker, squirrel	eagle, woodpecker, falcon
mural	sculpture, painting, portrait	painting, portrait, sculpture
tobacco	coffee, cigarette, corn	cigarette, cigar, corn
depth	size, bottom, meter	sea, underwater, level
chaos	anarchy, despair, demon	demon, anarchy, destruction

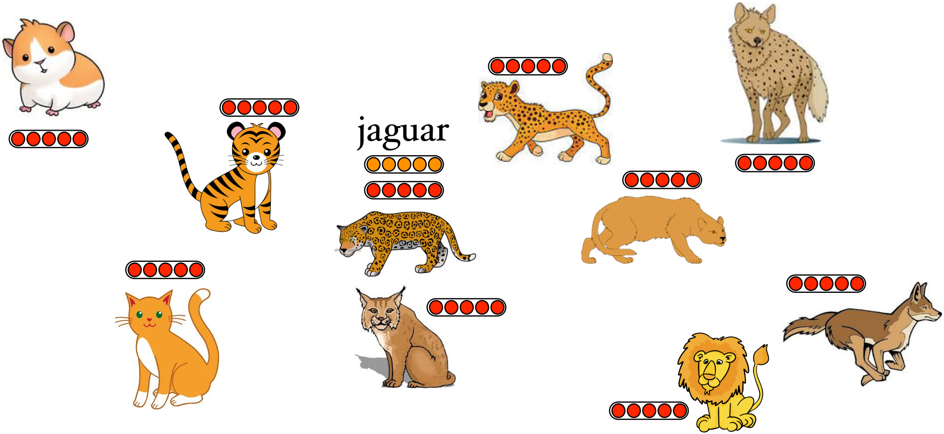
Out-of-the box 0-shot image retrieval with MSG

Training



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

	P@1	**P@10**	**P@20**	**P@50**
chance	<0.1	0.2	0.4	1.0
skip-gram/supervised cross-modal mapping	**2.3**	11.9	17.9	30.9
multimodal skip-gram/direct retrieval	2.0	**14.1**	**20.1**	**33.0**

Nearest visual neighbours of abstract words

freedom

theory

wrong

god

together

place

Subjects' significant preference for true neighbour over confounder:

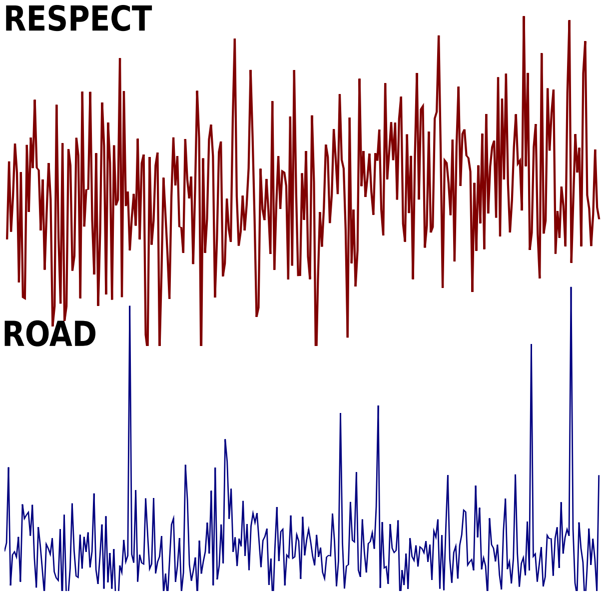
random level: 0%

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

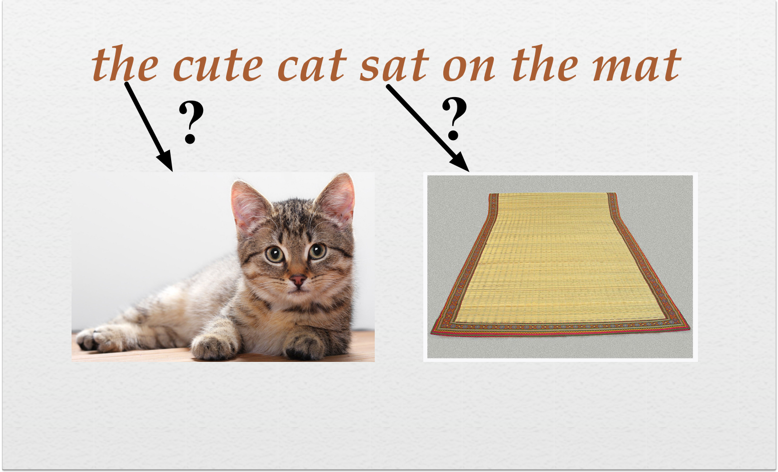
Real conversational data (ideally, child-directed speech)

```
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



Realistic word learning challenges for MSG

Learning from minimal exposure ("*fast mapping*")

The Frank corpus

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

```
*mot let me have that
%ref: RING
*mot ahhah whats 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

Model	*Best F*
MSG	.75
BEAGLE	.55
PMI	.53
Bayesian CSL	.54
(BEAGLE+PMI	.83)

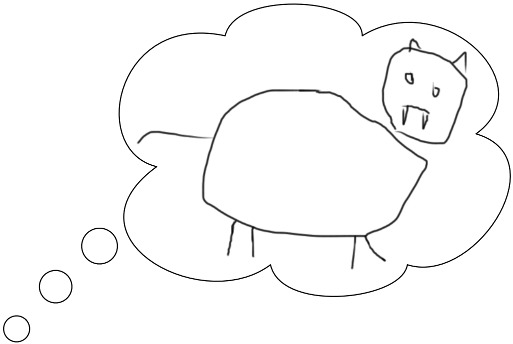
BEAGLE, PMI: Kievit-Kylar et al. CogSci 2013
Bayesian CSL: Frank et al. NIPS 2007

MSG object identification after a single exposure

word	*gold object*	*17 objects*	*5K objects*
bunny	bunny	bunny	hare
cows	cow	cow	heifer
duck	duck	hand	chronograph
duckie	duck	hand	chronograph
kitty	kitty	kitty	kitten
lambie	lamb	lamb	lamb
moocows	cow	pig	bison
rattle	rattle	hand	invader

And now for something (almost) completely different…

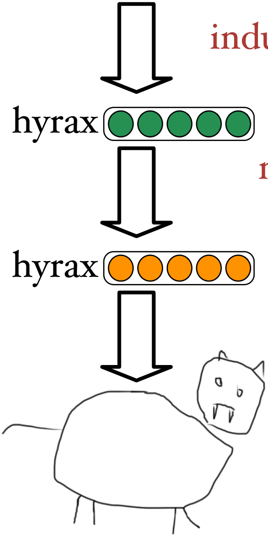
Imagining things you've never seen!



But there is another family member that is often forgotten: the **hyrax**! It might look a bit like a large guinea pig or rabbit with very short ears, but the **hyrax** is neither. Instead, the **hyrax** has similar teeth, toes, and skull structures to that of an elephant's. More importantly, the **hyrax** shares an ancestor with the elephant. The **hyrax**'s strong molars grind up tough vegetation, and two large incisor teeth grow out to be tiny tusks, just like an elephant's.

Generating pictures from word representations

But there is another family member that is often forgotten: the **hyrax**! It might look a bit like a large guinea pig or rabbit with very short ears, but the **hyrax** is neither. Instead, the **hyrax** has similar teeth, toes, and skull structures to that of an elephant's. More importantly, the **hyrax** shares an ancestor with the elephant. The **hyrax**'s strong molars grind up tough vegetation, and two large incisor teeth grow out to be tiny tusks, just like an elephant's.

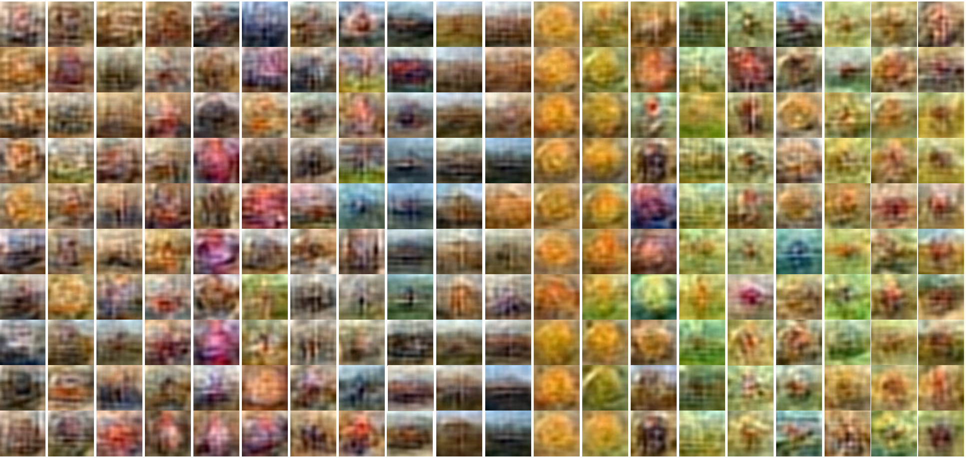


inducing word representations from text

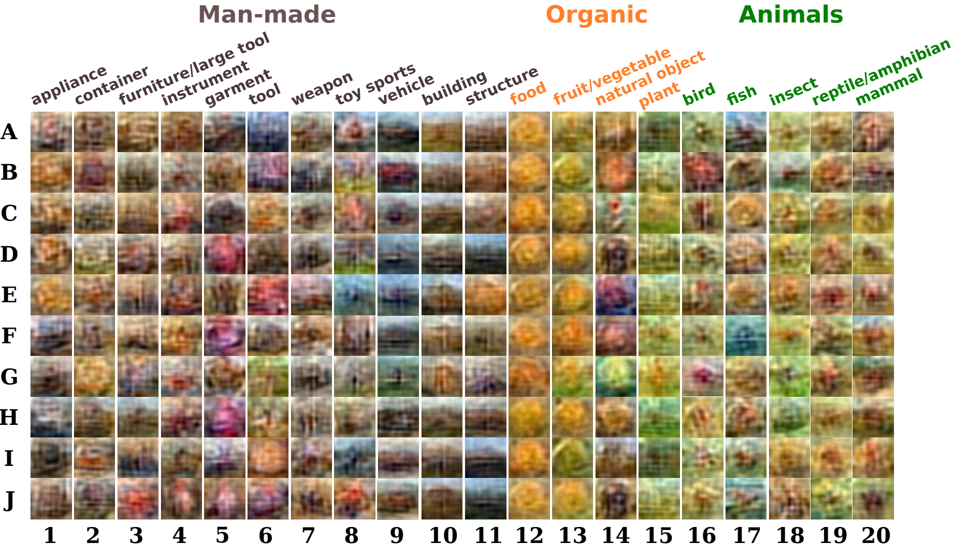
mapping onto high-level visual vectors

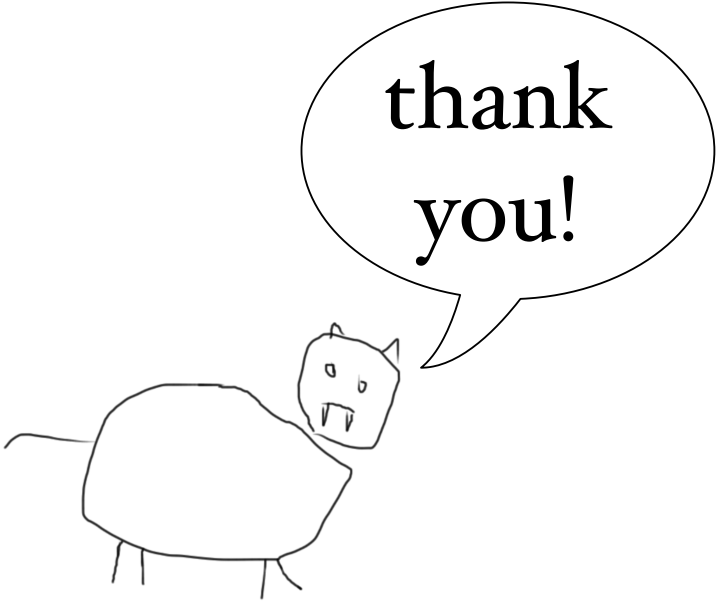
image generation from visual vectors

How word2vec sees the world



How word2vec sees the world




thank you!