Concepts, Composition, and Conversational Coordination

Semantic Competence for Situated Interaction

2nd Seminar: Concepts

David Schlangen
University of Potsdam, Germany

http://clp.ling.uni-potsdam.de

https://github.com/davidschlangen/cosine-paris

plan

the seminar series:

- intro: the problem & the approach
- concepts [Mon, Sep 23]
- composition [Mon, Sep 30]
- conversational coordination / dialogue [Mon, Oct 7]

functional *reprsnt.nal* *composition* *coordination*

Conceptual Apparatus

Inference

- word to word
- discourse resolution

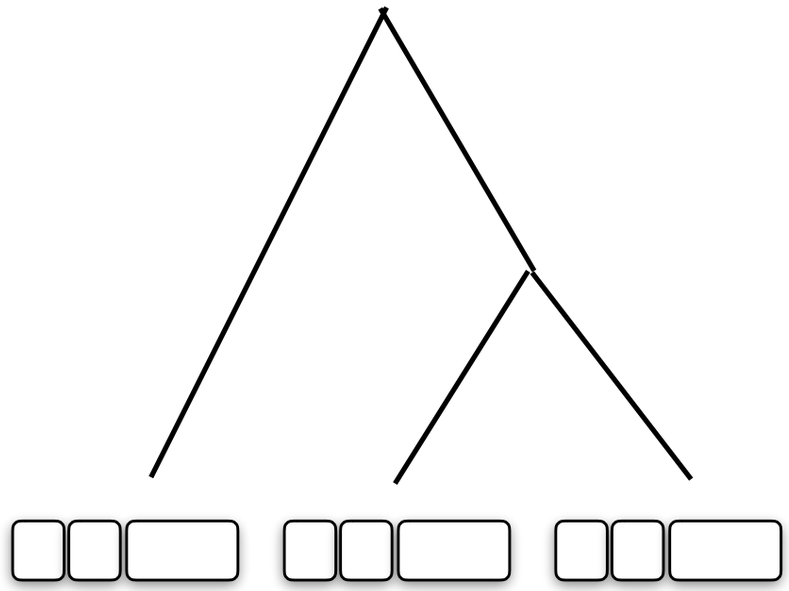
Reference

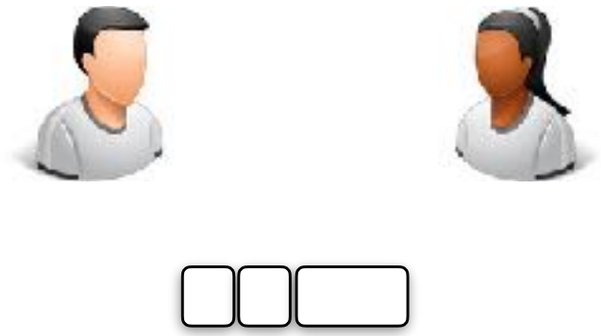
- word to world
- categori-sation
- naming / resolution

symbolic repr.

continuous repr.

classifiers on perceptual input





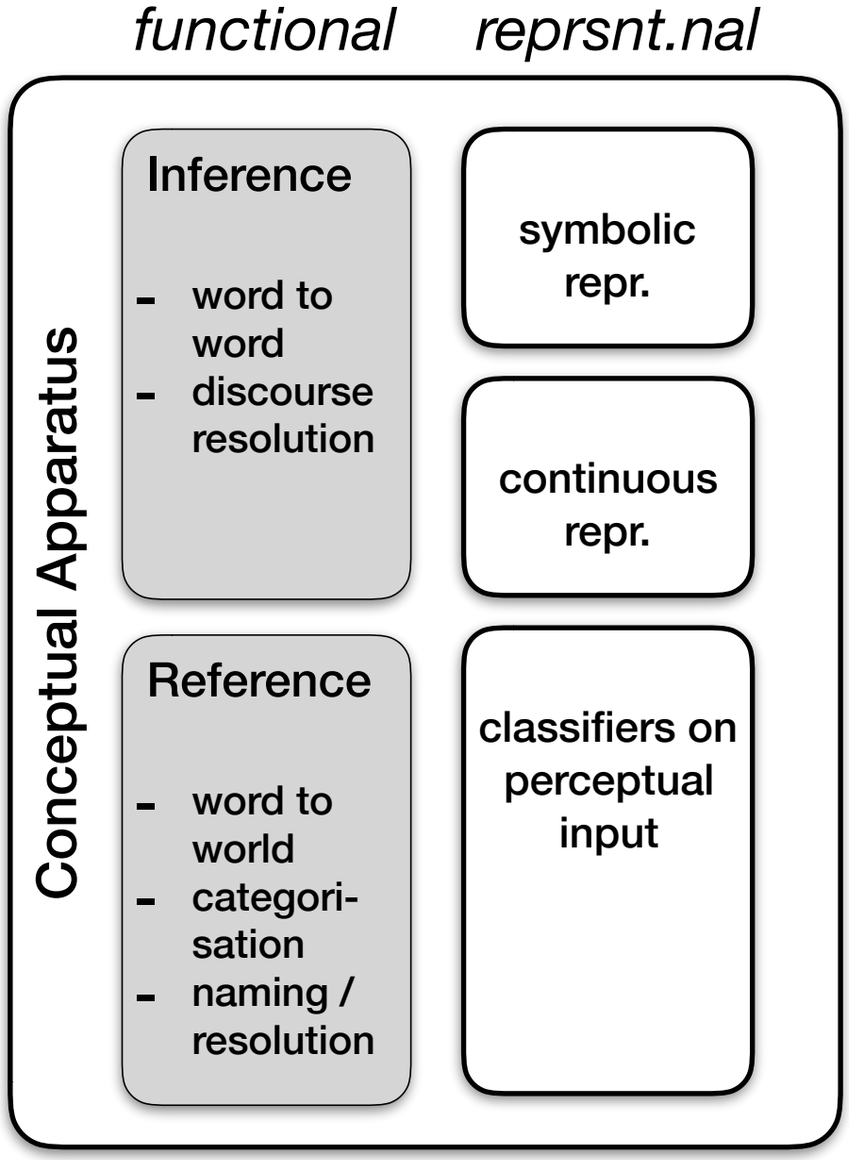
Look at the white dog!

We just saw a cute dog.

The cutest poodle ever!

Actually, that wasn't a poodle. It was too tall. It was a labradoodle.

- learning
  - incremental (within concept, within vocab)
  - fast
- implemented & tested on real data



- learning
  - incremental (within concept, within vocab)
  - fast
- implemented & tested on real data

- SAIAPR (2006 ff.): 20k images (Grubinger *et al.* 2006; Escalante *et al.* 2010), 200,000 referring expressions (Kazemzadeh *et al.* 2014)

- Flickr30k (2014 ff.): 30k images, 160,000 captions (Young *et al.* 2014)

- MSCOCO (2014 ff.): 300k images, 400,000 captions (Lin *et al.* 2014), 280,000 referring expressions (Yu *et al.* 2016)

- VisualGenome (2016 ff.): 100k images, 2e6 region descriptions (Krishna *et al.* 2016)

Can we create plausible learning and application situations from that data?

today

- data
  - what's in a language & vision corpus?
  - what's in the LV corpora that we've used?
- the "words as classifiers" model of referential concepts
  - basics: learning & simple application
  - what do they learn?
  - naming & factor graphs

today

- data
  - what's in a language & vision corpus?
  - what's in the LV corpora that we've used?
- the "words as classifiers" model of referential concepts
  - basics: learning & simple application
  - what do they learn?
  - naming & factor graphs

relations between corpus objects

expressions

(strings)

images

(pixels)

https://github.com/clp-research/sempix

relations between corpus objects

Two female tennis players shaking hands across the net.

$R_{paired\text{-}with}$

[COCO] (Lin *et al.* 2014)

relations between corpus objects

expressions
(strings)

images
(pixels)

words
sentences
discourses

referring expressions
region descriptions
captions
description paragraphs
q/a pairs
q/a sequences
dialogues

whole images
image regions

relations between corpus objects

Two female tennis players shaking hands across the net.

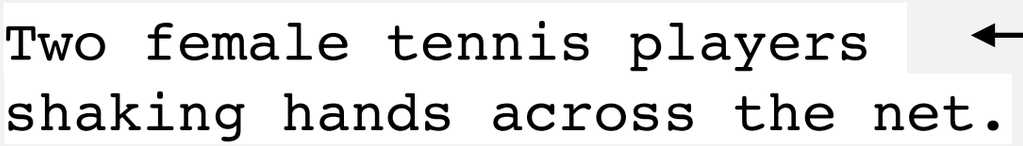
$R_{paired\text{-}with}$

describes

makes true

[COCO] (Lin *et al.* 2014)

images as *models*



$\models$

- reminder: in formal semantics, expressions are evaluated relative to *models*.

- M = < D, I >

- D: set of individuals

- I: interpretation function that maps non-logical constants to ([characteristic functions of] sets of) individuals

images as *models*

Two female tennis players shaking hands across the net.

$\models$ M = < D, I >, with:

D = {a, b, c}

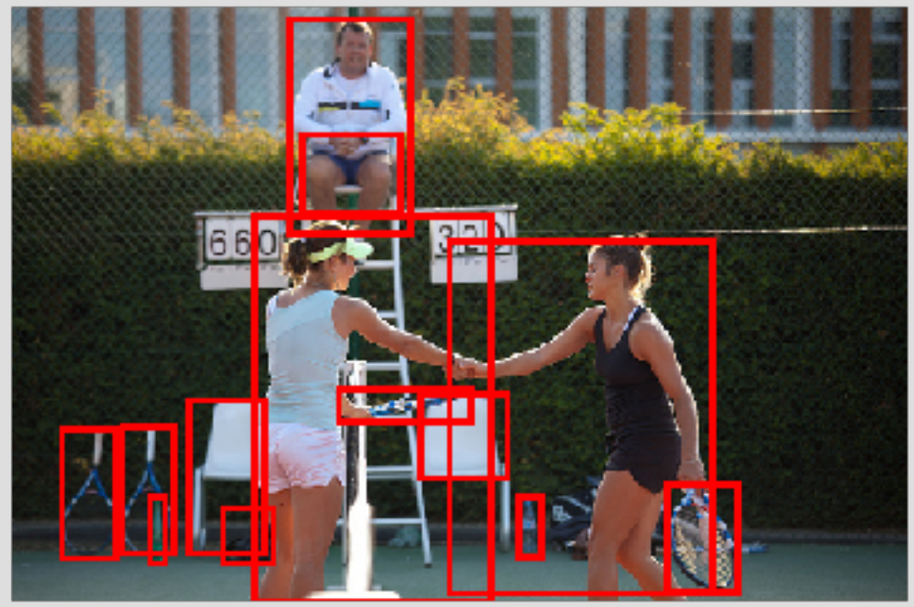
I(woman) = {a, b}
I(tennis_player) = {a, b}
I(shaking_hands) = {(a,b)}

images as *models*

Two female tennis players shaking hands across the net.

(Young *et al.* 2014)
(Hürliman & Bos, 2016)
(Schlangen *et al.*, 2016)

images as *models*

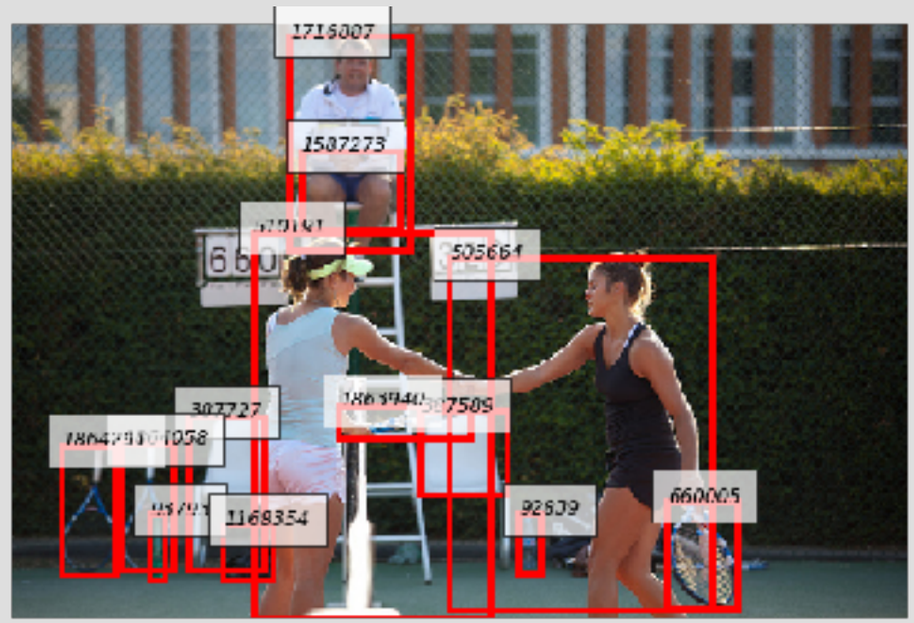


$\models$

Two female tennis players shaking hands across the net.

images as *models*

Two female tennis players shaking hands across the net.



images as *models*

Two female tennis players shaking hands across the net.

$$D = \{o_{92839}, o_{93793}, o_{387589}, o_{387727}, o_{505664}, o_{510191}, o_{660005}, o_{1168354}, o_{1587273}, o_{1716887}, o_{1863940}, o_{1864058}, o_{1864291}\}$$

images as *models*

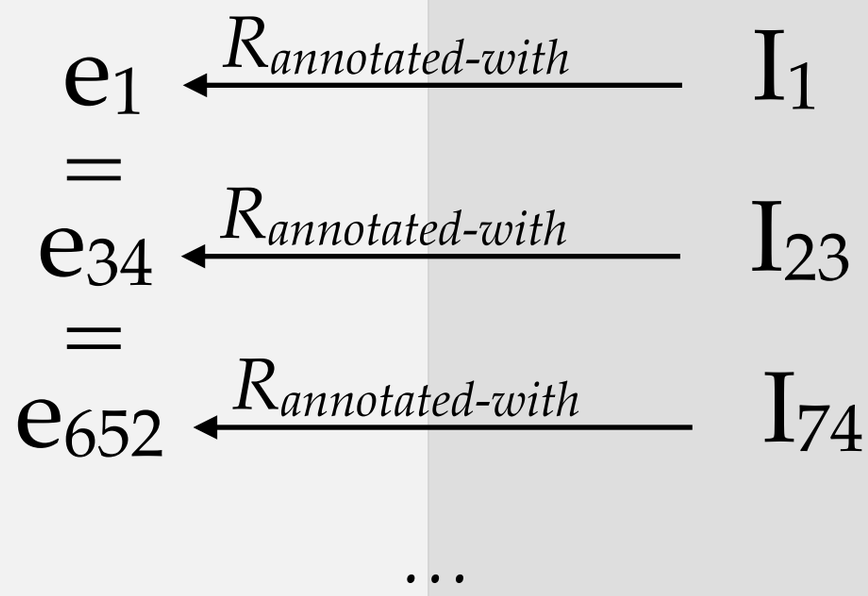
Two female tennis players shaking hands across the net.

$$D = \{o_{92839}, o_{93793}, o_{387589}, o_{387727}, o_{505664}, o_{510191}, o_{660005},$$
$$o_{1168354}, o_{1587273}, o_{1716887}, o_{1863940}, o_{1864058}, o_{1864291}\}$$

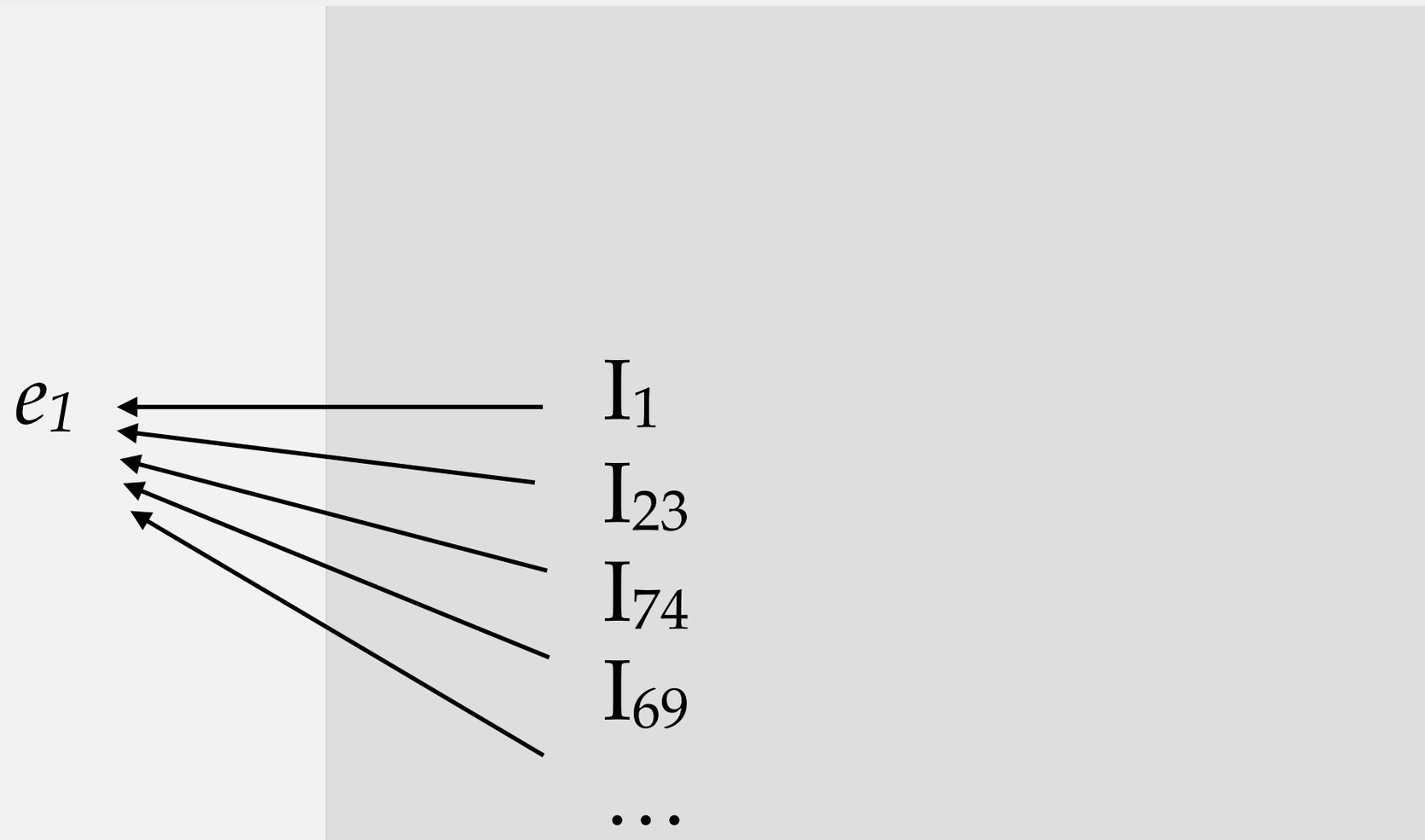
$$I(\textit{player}) = \{o_{505664}, o_{510191}\}$$

relations between corpus objects

relations between corpus objects



annotates same object (type)

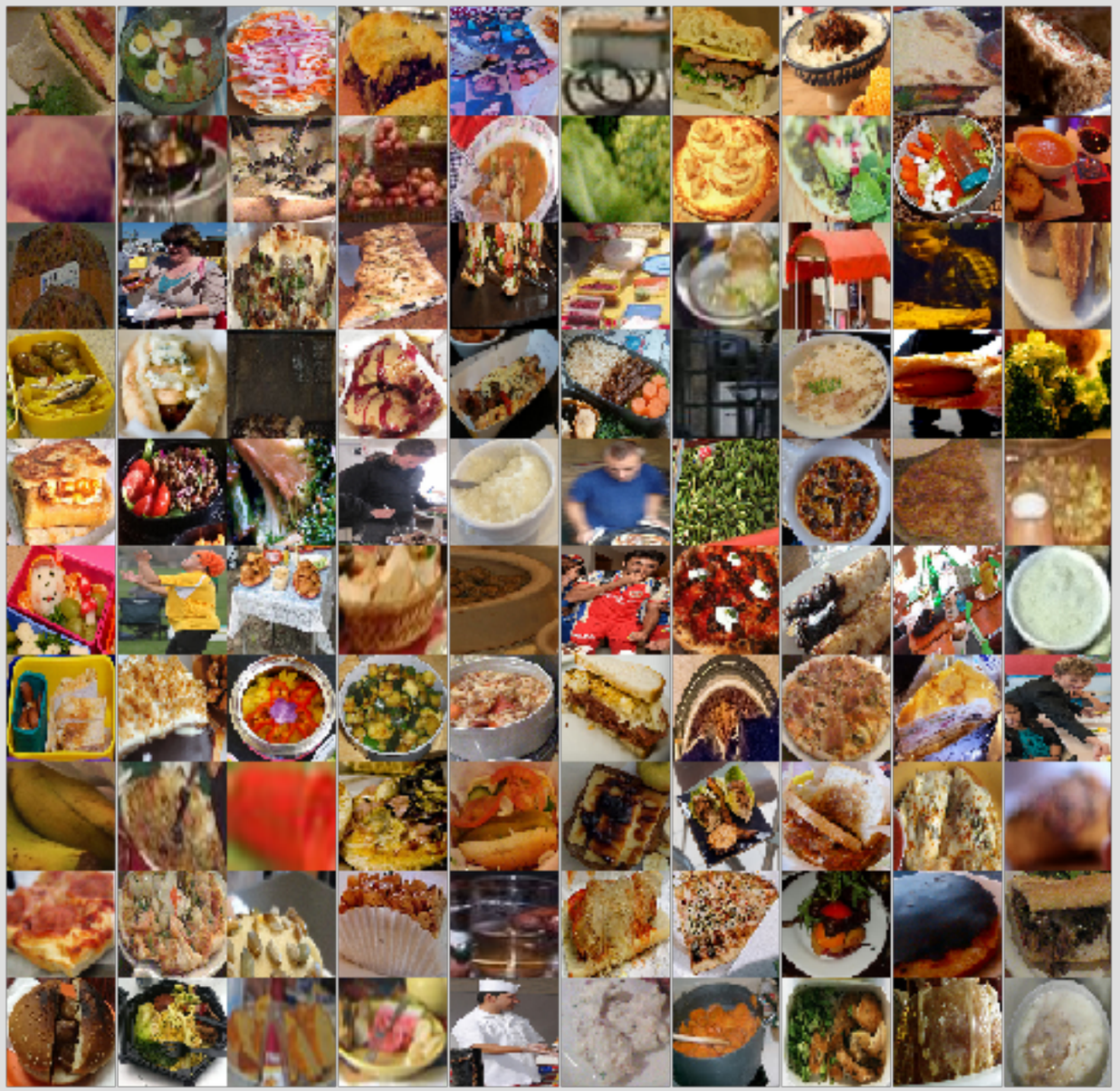


visual denotations

dog

visual denotations

food

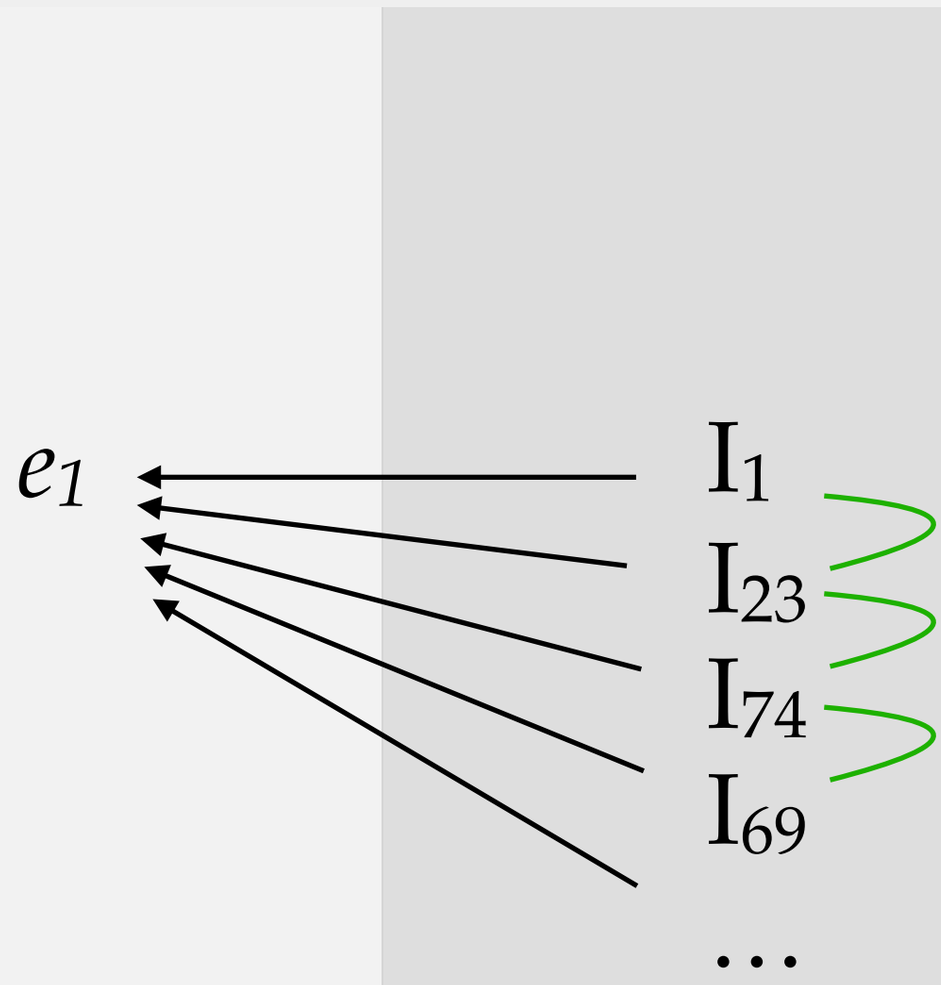


visual denotations

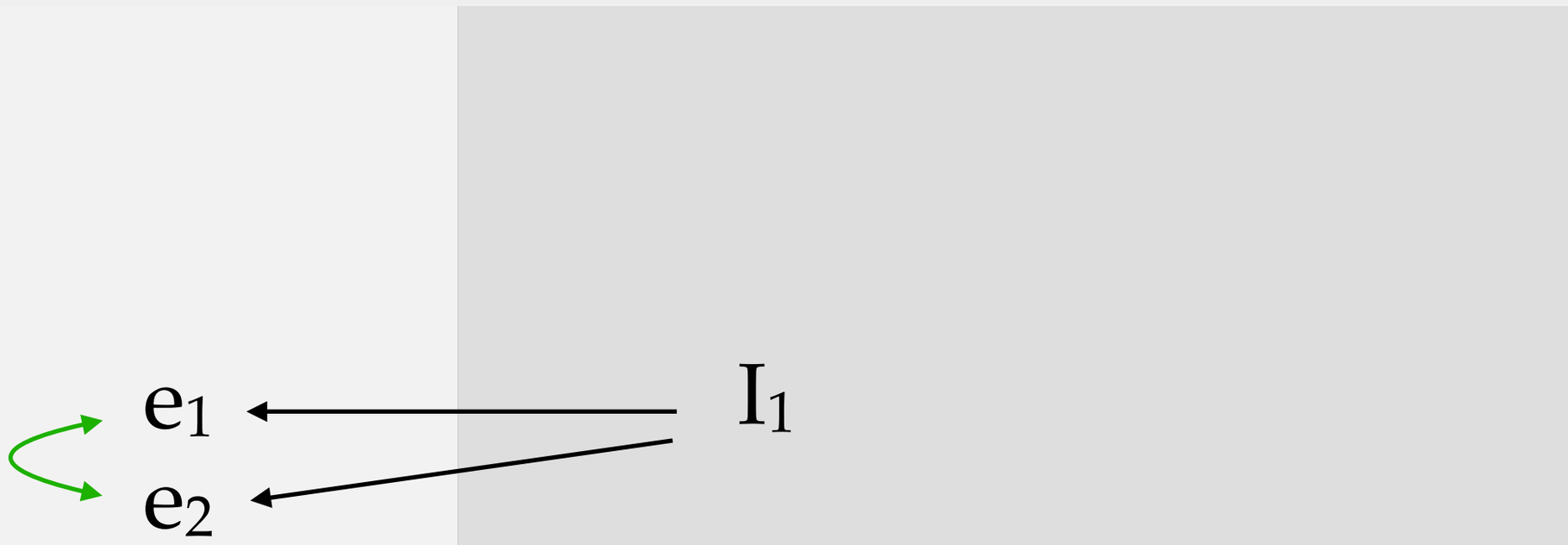
green

- maybe do live demo of notebook here?

annotates same object (type)



annotates same object



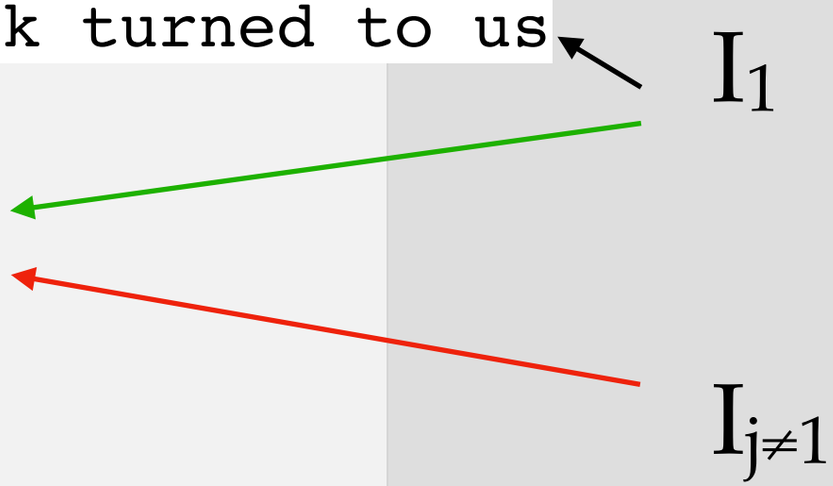
guy on left with back turned to us

left kid

A: guy on left with back turned to us

B: hu?

A: left kid

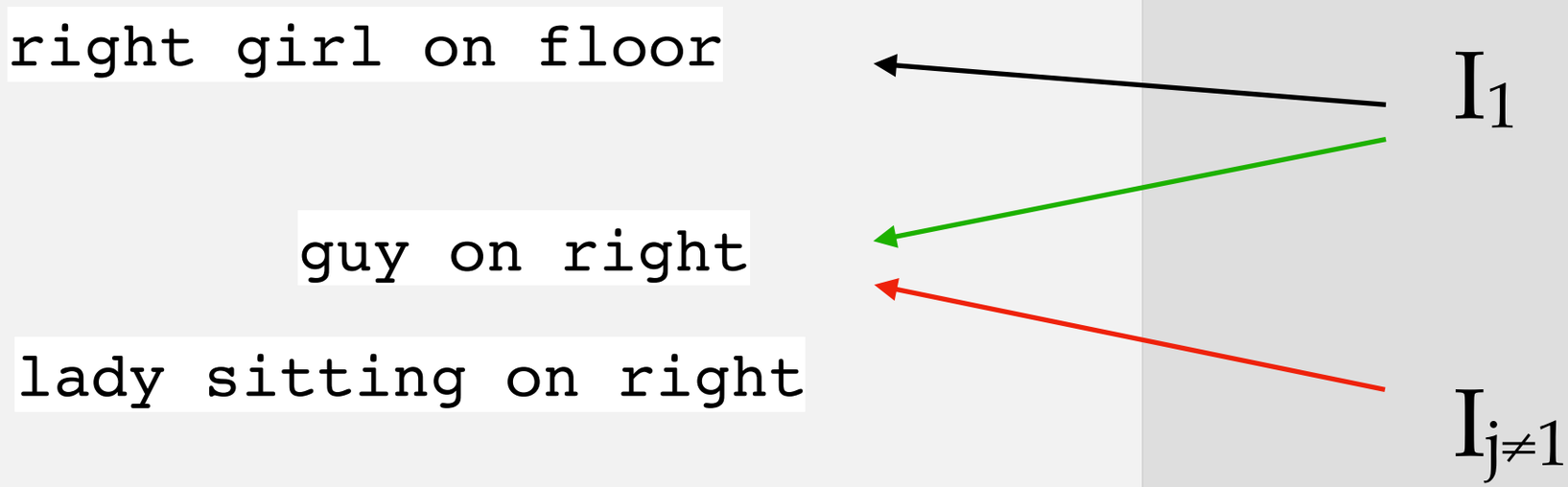


guy on left with back turned to us

left kid

left cop

situational entailment



types of objects

caption
region description
referring expression
paragraph

$R_{annotated\text{-}with}$

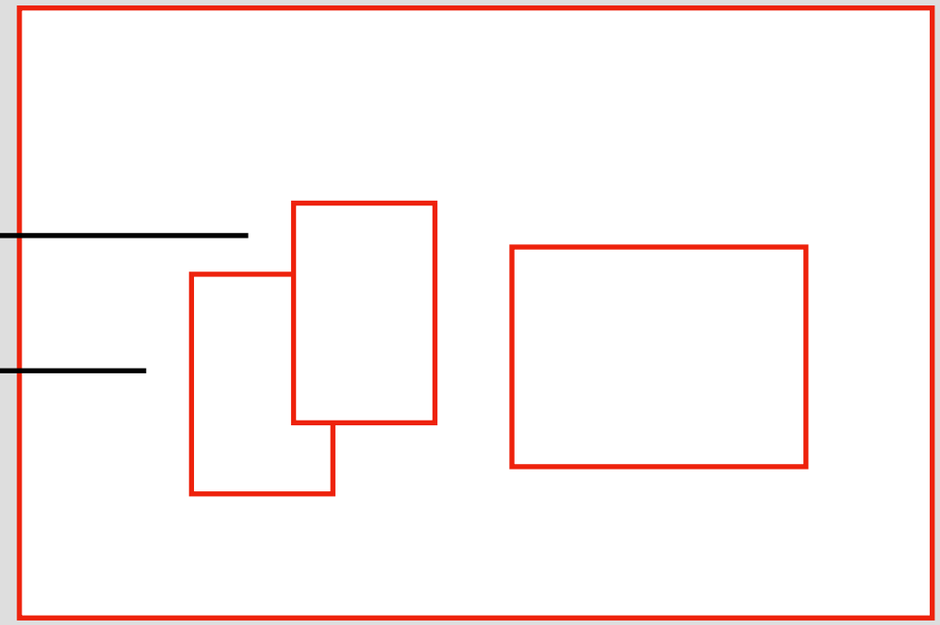


image
image region

A bronze car

A car parked outside

A person skiing downhill

image / image similarity

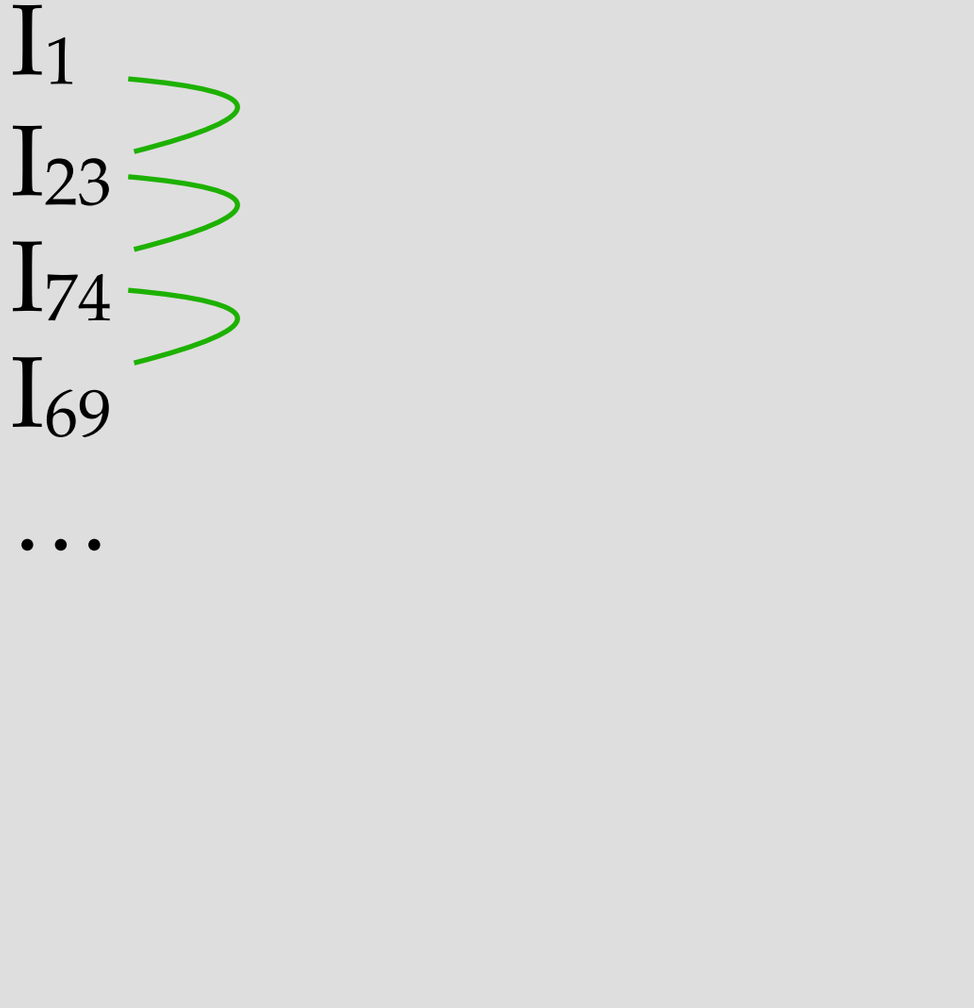


image / image similarity

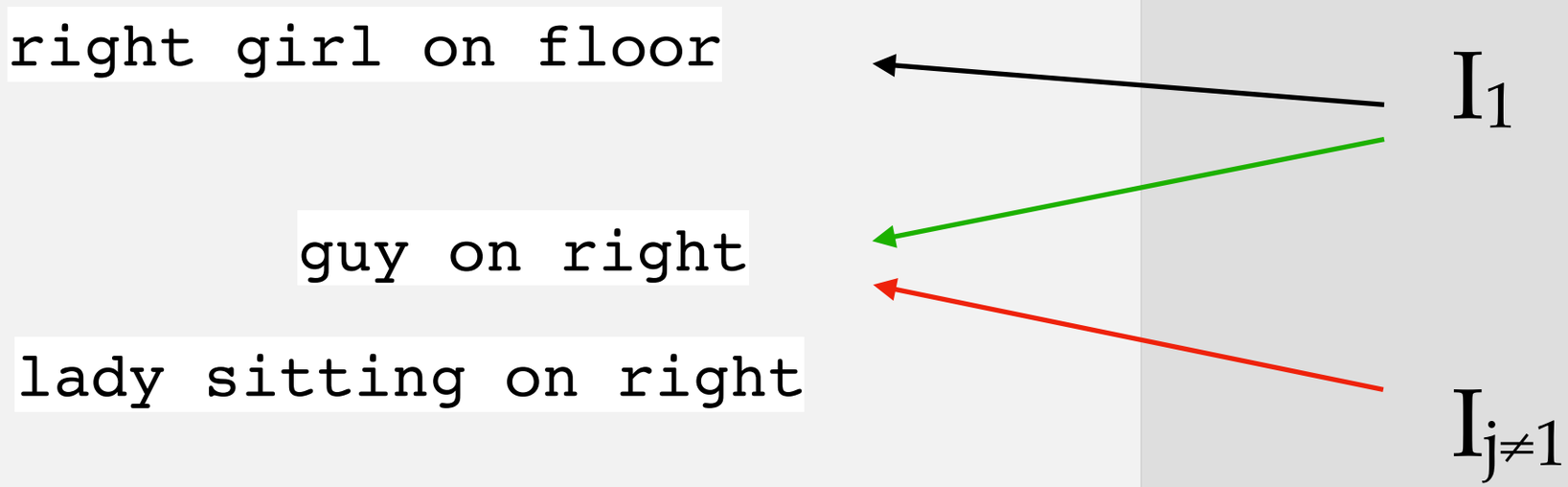
situational entailment

A woman with a painted face riding a skateboard indoors

A woman with face paint on standing on a skateboard

There are men who are skateboarding down the trail.

situational entailment



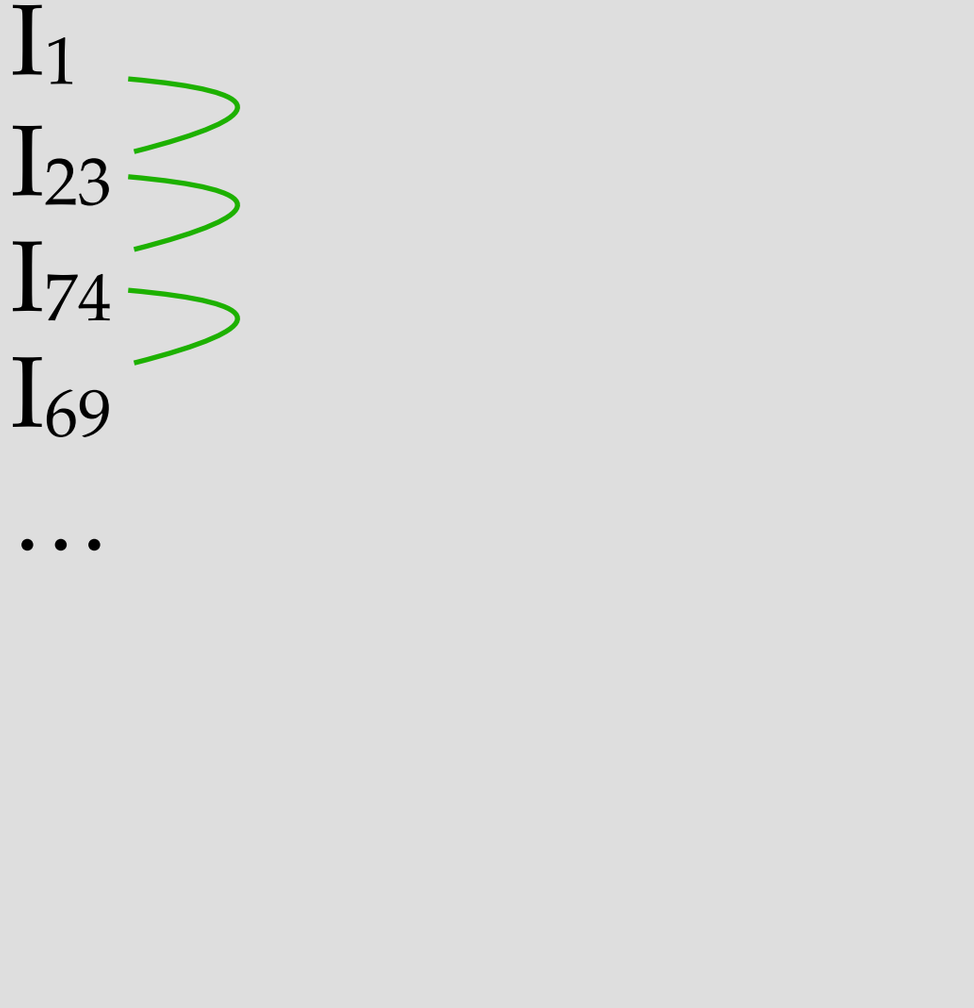
- randomly sampled 60 pairs (balanced pos / neg)
- asked 3 workers each on AMT “*text 2 refers to same object as text 1*”, agreement on 4-step Likert scale
- accuracy (majority): 0.68

(Schlangen, IWCS 2019)

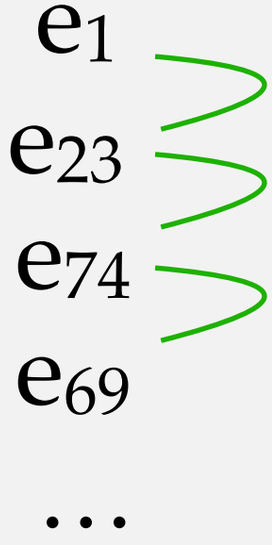
situational entailment

- referring expression / referring expression: 0.68
- caption / caption: 0.63
- caption / objects (“there is a”): 0.58
- caption / region description: 0.6
- caption / paragraph: 0.6

image / image similarity

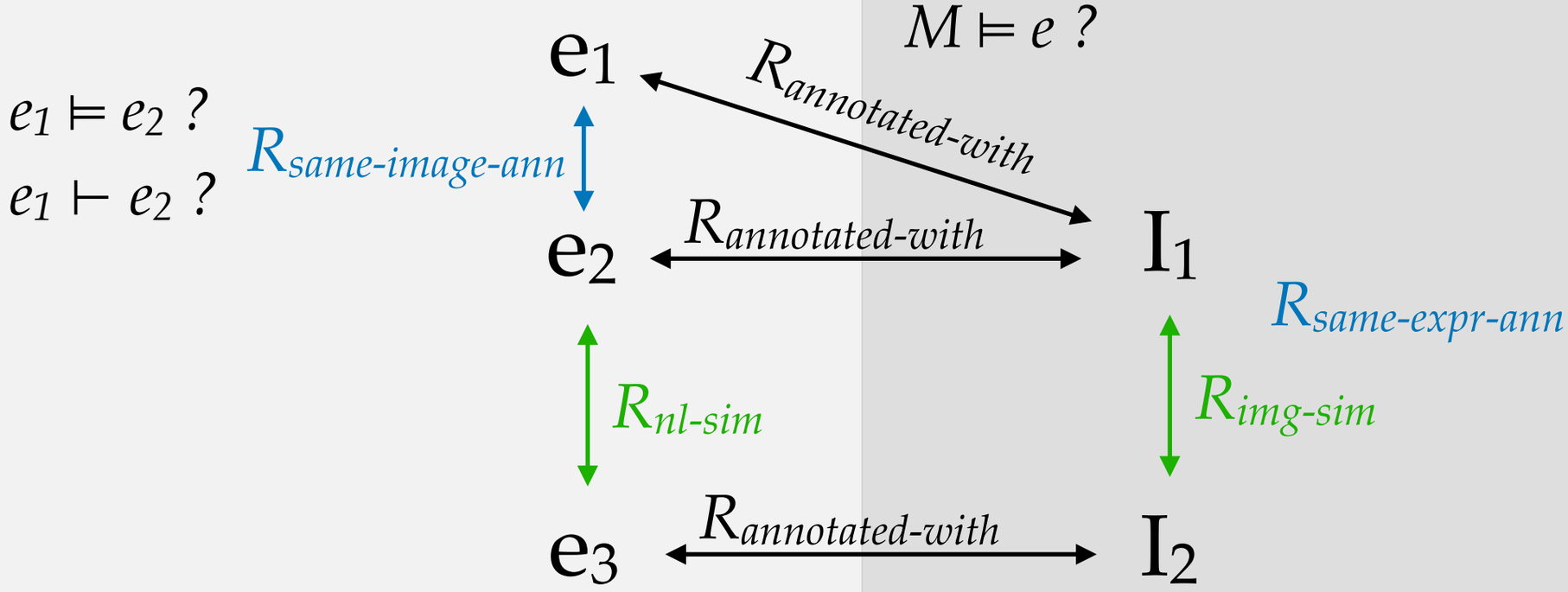


expression /
expression similarity



(Cer *et al.* 2018)

corpus object relations in the corpora



Besides the *primary* relations in the corpora (captioning, referring) we can derive further relations between expressions & images, expressions & expressions, images & images.

“annotated with”: captioning / *describe*

Instructions:

- Describe all the important parts of the scene.
- Do not start the sentences with "There is".
- Do not describe unimportant details.
- Do not describe things that might have happened in the future or past.
- Do not describe what a person might say.
- Do not give people proper names.
- The sentence should contain at least 8 words.

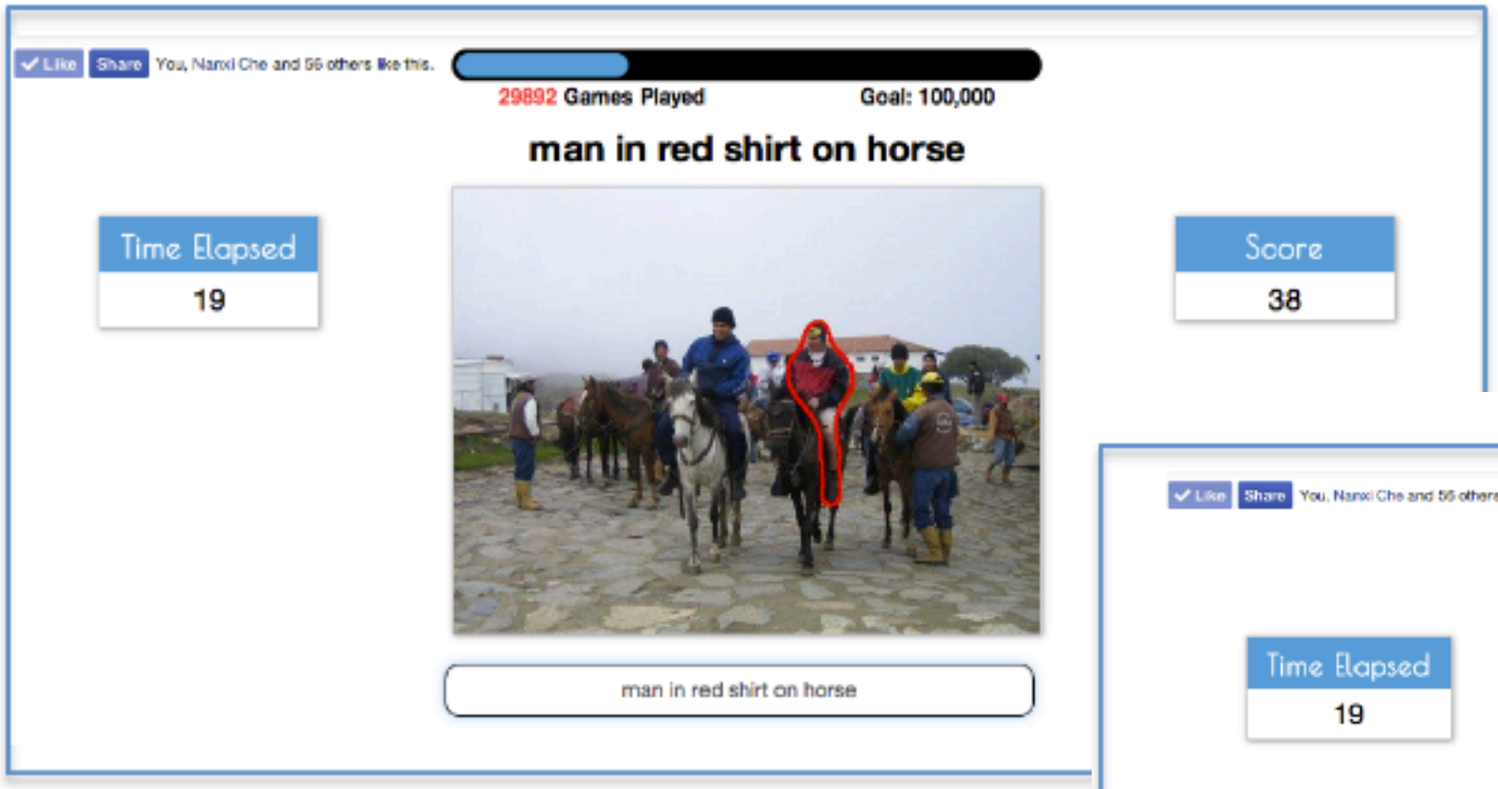
Please describe the image:

Enter description here

prev next

(Chen *et al.* 2015)

“annotated with”: ref-exp / *single out*



(Kazemzadeh *et al.* 2014)

- SAIAPR (2006 ff.): 20k images (Grubinger *et al.* 2006; Escalante *et al.* 2010), 200,000 referring expressions (Kazemzadeh *et al.* 2014)

- Flickr30k (2014 ff.): 30k images, 160,000 captions (Young *et al.* 2014)

- MSCOCO (2014 ff.): 300k images, 400,000 captions (Lin *et al.* 2014), 280,000 referring expressions (Yu *et al.* 2016)

- VisualGenome (2016 ff.): 100k images, 2e6 region descriptions (Krishna *et al.* 2016)

Can we create plausible learning and application situations from that data?

… bringing in agents

- take (e, I) pairs as *observations* of agents:

`Two female tennis players shaking hands across the net.`

$R_{paired\text{-}with}$

… bringing in agents

- take (e, I) pairs as *observations* of agents:

How nice! Two female tennis players shaking hands across the net.

… bringing in agents

- An interpreted observation of an utterance of *e*:

$$o = \langle e, i, \delta, R_\delta, \omega, R_\omega \rangle$$

Inspired by (Erk 2013)

- i: spatio-temporal parameters, e.g. speaker, time, place
- δ: discourse context
- R_δ : relation of utterance to discourse context
- ω: visual context
- R_ω: relation of utterance to visual context (e.g. *describes, refers to)*

… bringing in agents

- linguistic history of agent A is set of all observations: O^A
- accessor functions, e.g. expr(o) is the utterance type
- all observations of *e* by A: $O_e^A = \{o \in O^A \mid expr(o) = e\}$
- if word w in expr(o), then o is in O_w
- what can we learn about w from O_w ?
- how do $O_{w,\ t<now}$ determine [[w]] ?
- how does a given o_w update [[w]]?

today

- data
  - what's in a language & vision corpus?
  - what's in the LV corpora that we've used?
- the "words as classifiers" model of referential concepts
  - basics: learning & simple application
  - what do they learn?
  - naming & factor graphs

today

- data
  - what's in a language & vision corpus?
  - what's in the LV corpora that we've used?
- the "words as classifiers" model of referential concepts
  - basics: learning & simple application
  - what do they learn?
  - naming & factor graphs

functional *reprsnt.nal*

Conceptual Apparatus

Reference

- word to world
- categori-sation
- naming / resolution

classifiers on perceptual input

Look at the white dog!

$$[[dog]]^{D,I} = I[f_{dog}](D) = \{o \mid \mathbb{1}_{dog}(o) = 1, o \in D\}$$

$$[[dog]]^{D} = \{ (o, f_{dog}(o)) \}$$

what kind of function is this? (what is the range?)

where do we get this function?

how do we present image object to function?

what kind of set is this?

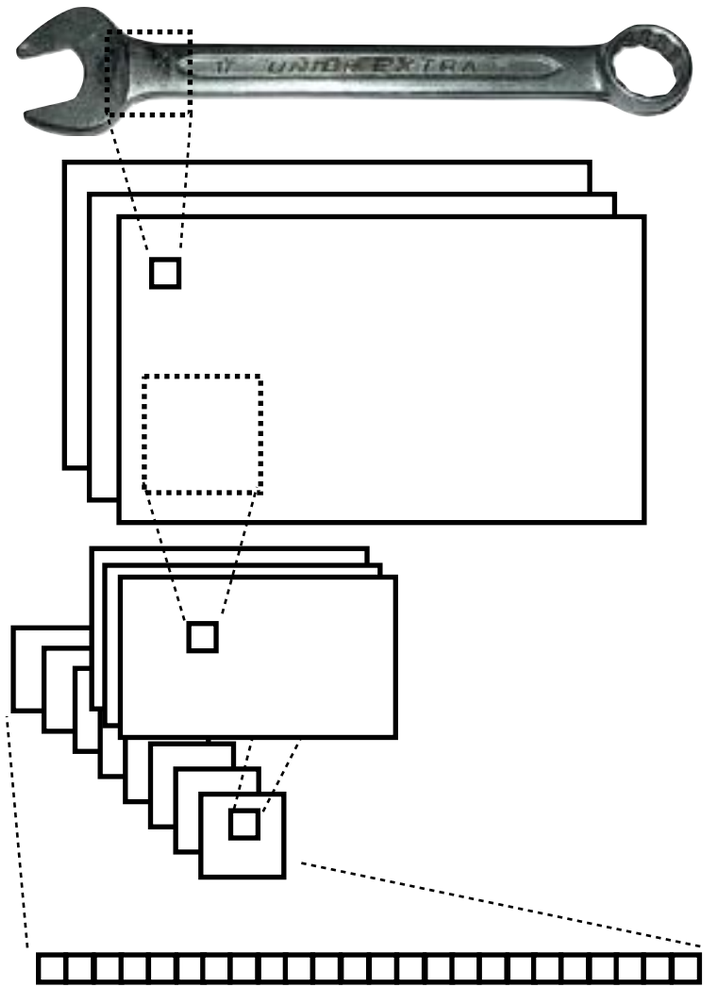
$$[[\text{wrench}]]^{D} = \{ (o, f_{\text{wrench}}(o)) \}$$

what kind of function is this? (what is the range?)

where do we get this function?

how do we present image object to function?

what kind of set is this?



convolutional NN, trained on ImageNet 1000, fully connected layer.

In (Schlangen *et al.* 2016), GoogLeNet (Szegedy *et al.* 2015), here ResNet 50 (He *et al.* 2015)

2048 CNN features
+ 7 positional features

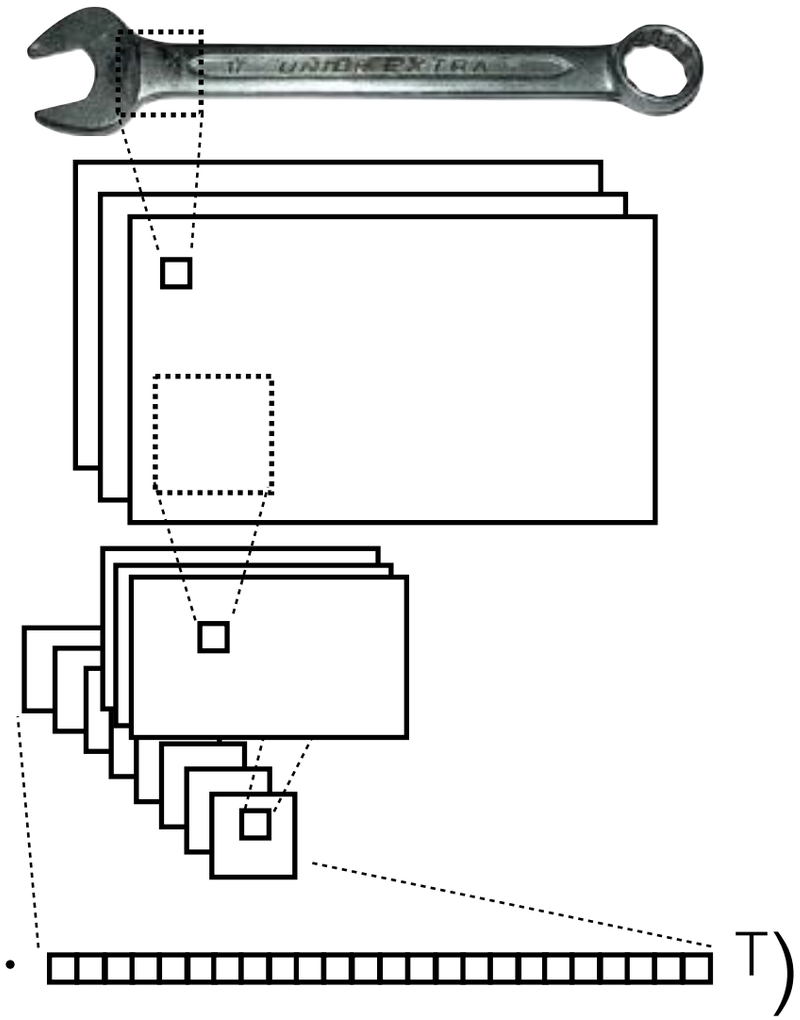
$$[[\text{wrench}]]^{D} = \{ (o, f_{\text{wrench}}(o)) \}$$

what kind of function is this? (what is the range?)

where do we get this function?

what kind of set is this?

how do we present image object to function?



Logistic regression, maps input (image object) to value in [0, 1].

convolutional NN, trained on ImageNet 1000, fully connected layer.

In (Schlangen *et al.* 2016), GoogLeNet (Szegedy *et al.* 2015), here ResNet 50 (He *et al.* 2015)

2048 CNN features
+ 7 positional features

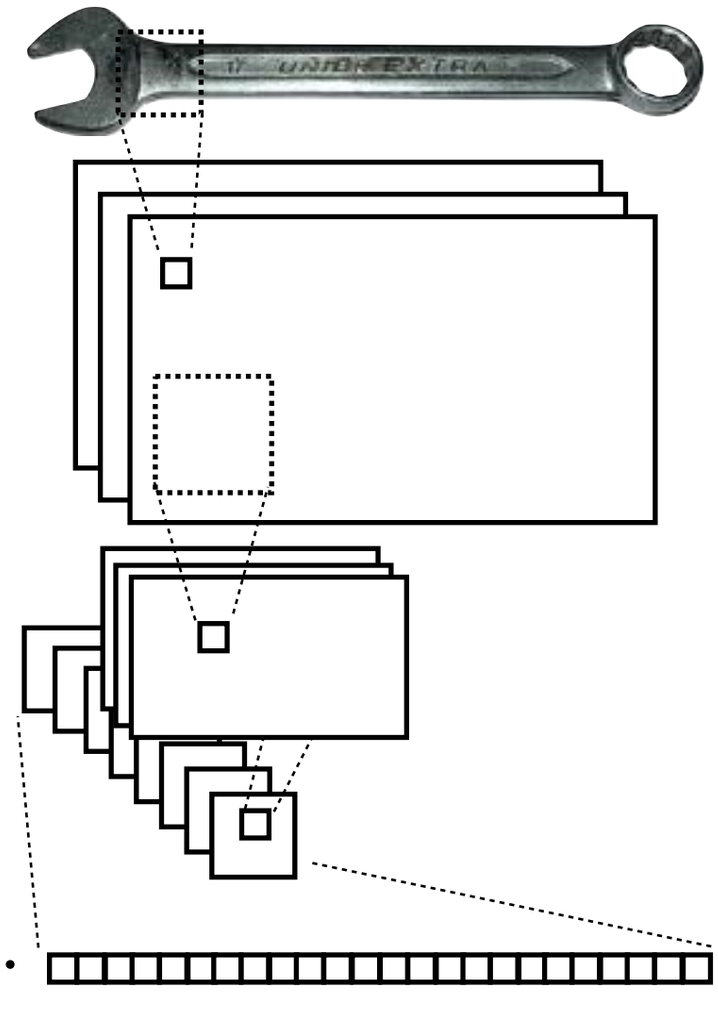
$$[[\text{wrench}]]^{D} = \{ (o, f_{\text{wrench}}(o)) \}$$

what kind of function is this? (what is the range?)

where do we get this function?

what kind of set is this?

how do we present image object to function?



$$\sigma(\Theta \cdot \quad \top)$$

[0,1]

Logistic regression, maps input (image object) to value in [0, 1].

cross entropy loss function, SGD, L1-regulated

convolutional NN, trained on ImageNet 1000, fully connected layer.

In (Schlangen *et al.* 2016), GoogLeNet (Szegedy *et al.* 2015), here ResNet 50 (He *et al.* 2015)

2048 CNN features
+ 7 positional features

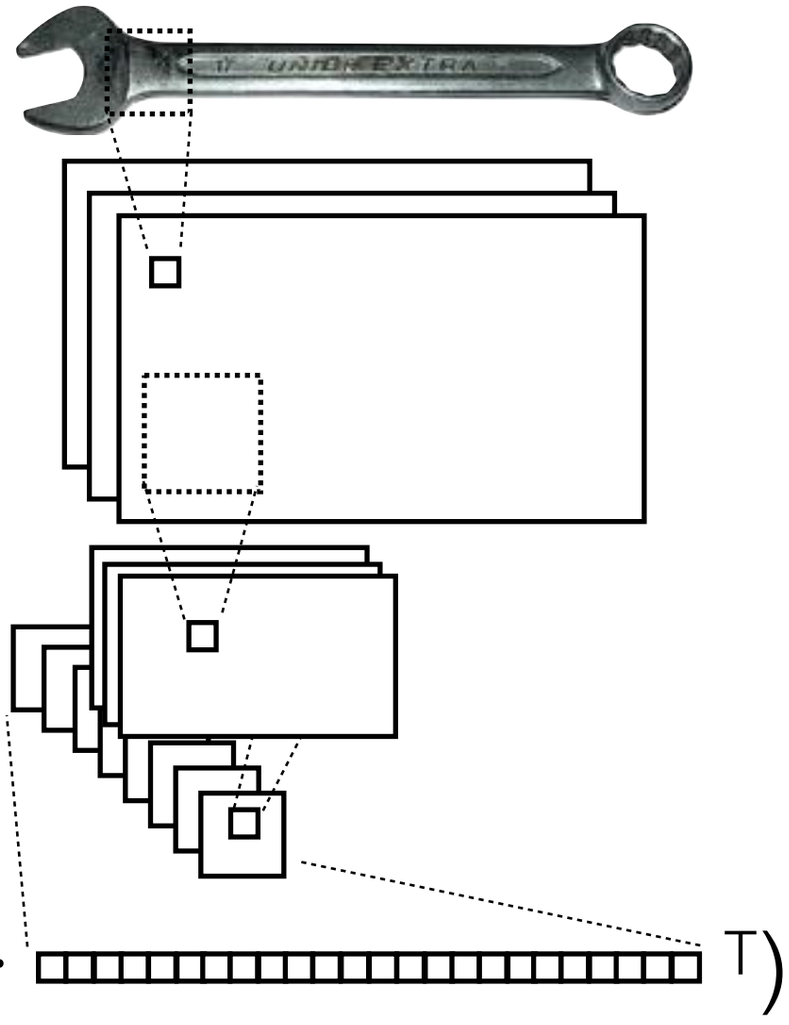
$$[[\text{wrench}]]^D = \{ (o, f_{\text{wrench}}(o)) \}$$

what kind of function is this? (what is the range?)

where do we get this function?

how do we present image object to function?

what kind of set is this?

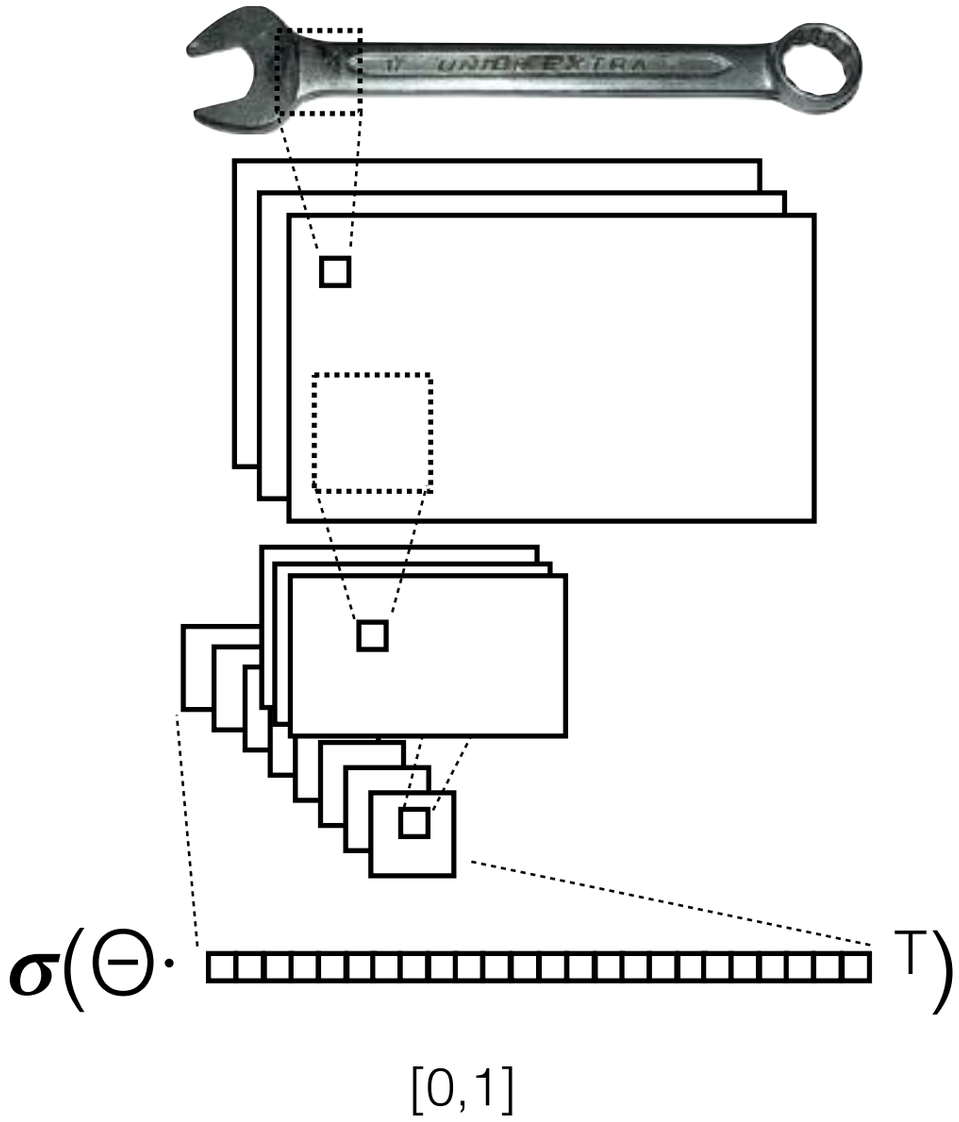


$$\sigma(\Theta \cdot \quad \top)$$

$[[\text{wrench}]]^D = \{ (o_1, 0.4),$ $(o_2, 0.98),$ $(o_3, 0.2) \}$

Logistic regression, maps input (image object) to value in [0, 1].

cross entropy loss function, SGD, L1-regulated



convolutional NN, trained on ImageNet 1000, fully connected layer.

In (Schlangen *et al.* 2016), GoogLeNet (Szegedy *et al.* 2015), here ResNet 50 (He *et al.* 2015)

2048 CNN features
+ 7 positional features

- ✅ learning
  - incremental
    - ✅ within concept
    - ✅ within vocabulary
  - ❓ fast
- ❓ implemented & tested on real data

learning

- from *ostensive definition* / labelled examples
- associative learning vs learning from interpreted observations.

“When they (my elders) named some object, and accordingly moved towards something, I saw this and I grasped that the thing was called by the sound they uttered when they meant to point it out. Their intention was shewn by their bodily movements, as it were the natural language of all peoples […]”

St. Augustine of Hippo (400)

[to teach children,] people ordinarily show them the thing whereof they would have them have the idea; and the repeat to them the name that stands for it: as *white, milk, sugar, cat, dog.*

John Locke (1690)

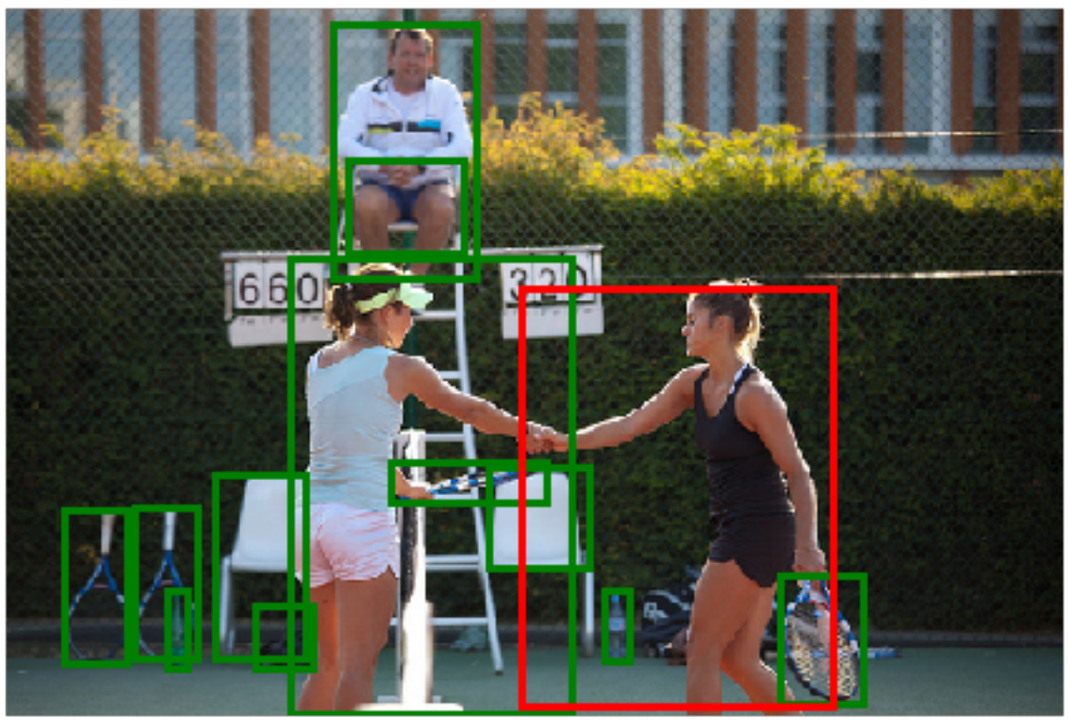
learning

- from *ostensive definition* / labelled examples
- associative learning vs learning from interpreted observations.
- Bloom (2000), E. Clark: learning from *intentions*
- $o = \langle e, i, \delta, R_\delta, \omega, R_\omega \rangle$, with R_ω = *refers-to*, pointing into ω

training data

- referring expressions, collected with ReferIt Game:
  - for SAIAPR, 120k. avg length: 3.43 token, 87% NP, 7% S
  - for MSCOCO, 140k rex (refcoco), 3.5 token
    + 140k (recoco+) w/ positional words, 3.53 token
  - collected non-interactively (Mao *et al.* 2016), 100k, 8.31 tk

```
refcoco_refdf :
-  lady in black on right
-  girl in black
-  woman in black
refcocoplus_refdf :
-  black shirt
-  girl in black
-  player in black
grex_refdf :
-  woman in black tank top and shorts holding
tennis racket
-  woman in black outfit shacking other tennis
player hand
```



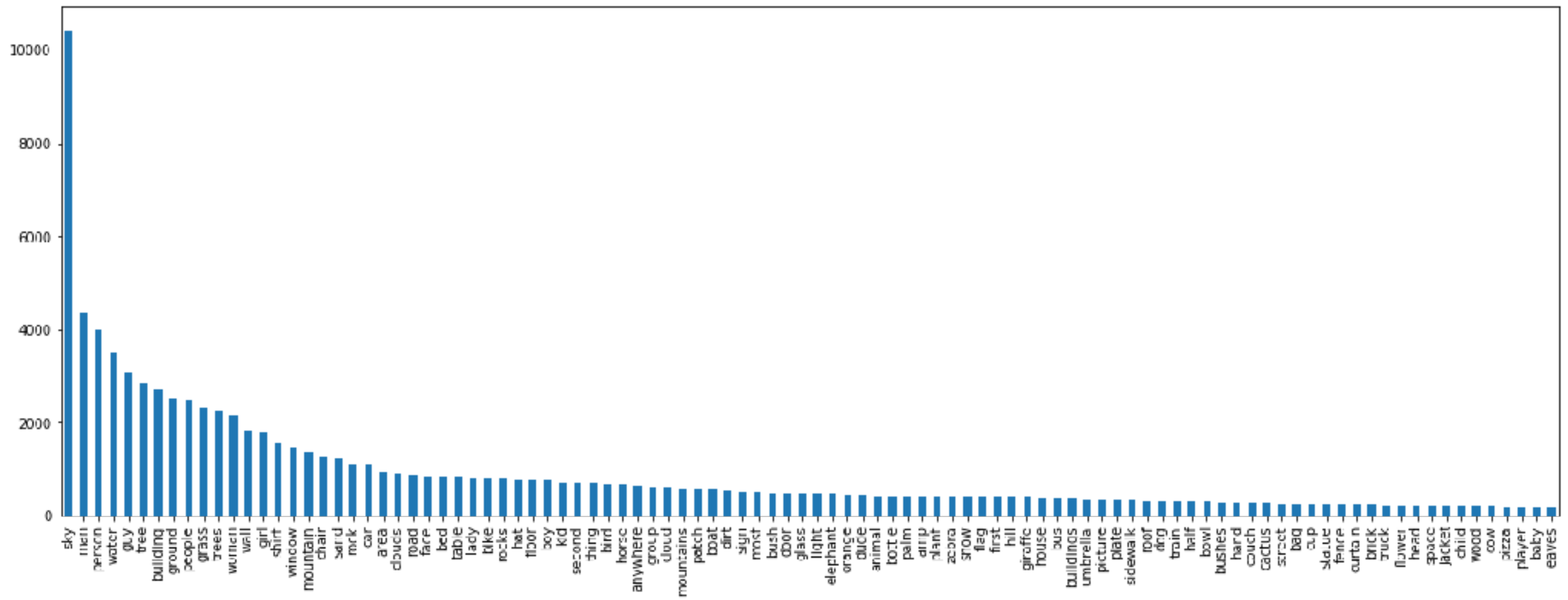
training data

- referring expressions, collected with ReferIt Game:
  - for SAIAPR, 120k. avg length: 3.43 token, 87% NP, 7% S
  - for MSCOCO, 140k rex (refcoco), 3.5 token
    + 140k (recoco+) w/ positional words, 3.53 token
- collected non-interactively (Mao *et al.* 2016), 100k, 8.31 tk
- structures (using Berkeley parser, [Kitaev & Klein 2018]):

	NP	S	Other
referit	87%	7%	6%
refcoco	83%	9%	8%
refcoco+	84%	12%	4%
grex	85%	14%	1%

training data

- vocabulary (head nouns, all ReferIt game corpora combined, validation split)



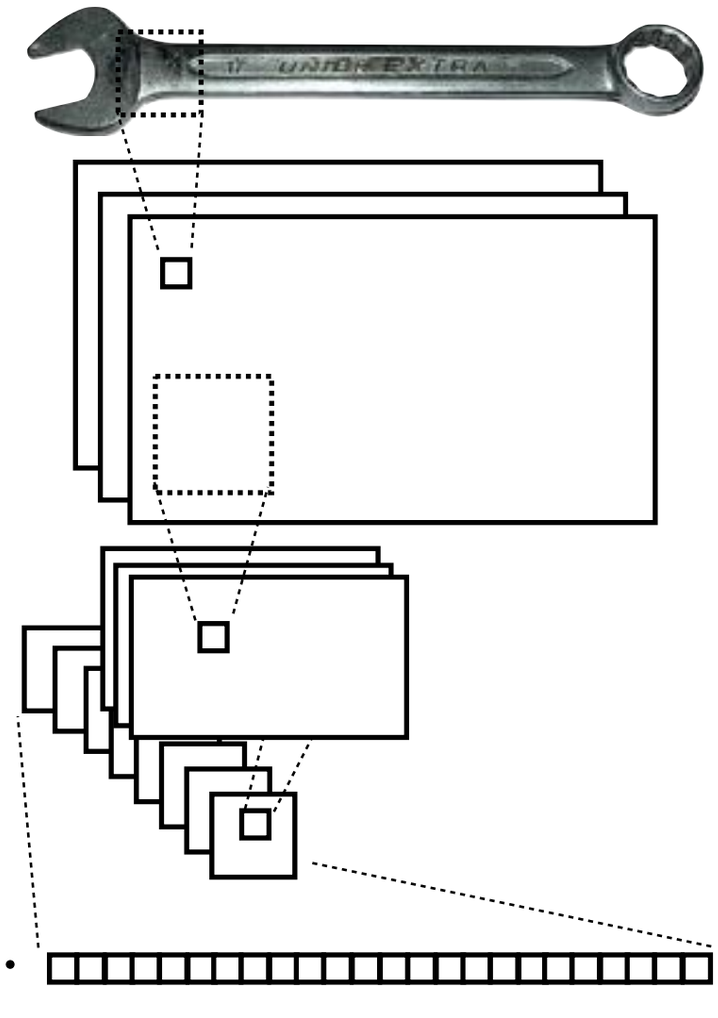
$$[[\text{wrench}]]^{D} = \{ (o, f_{\text{wrench}}(o)) \}$$

what kind of function is this? (what is the range?)

where do we get this function?

what kind of set is this?

how do we present image object to function?



$$\sigma(\Theta \cdot \quad \top)$$

[0,1]

Logistic regression, maps input (image object) to value in [0, 1].

cross entropy loss function, SGD, L1-regulated

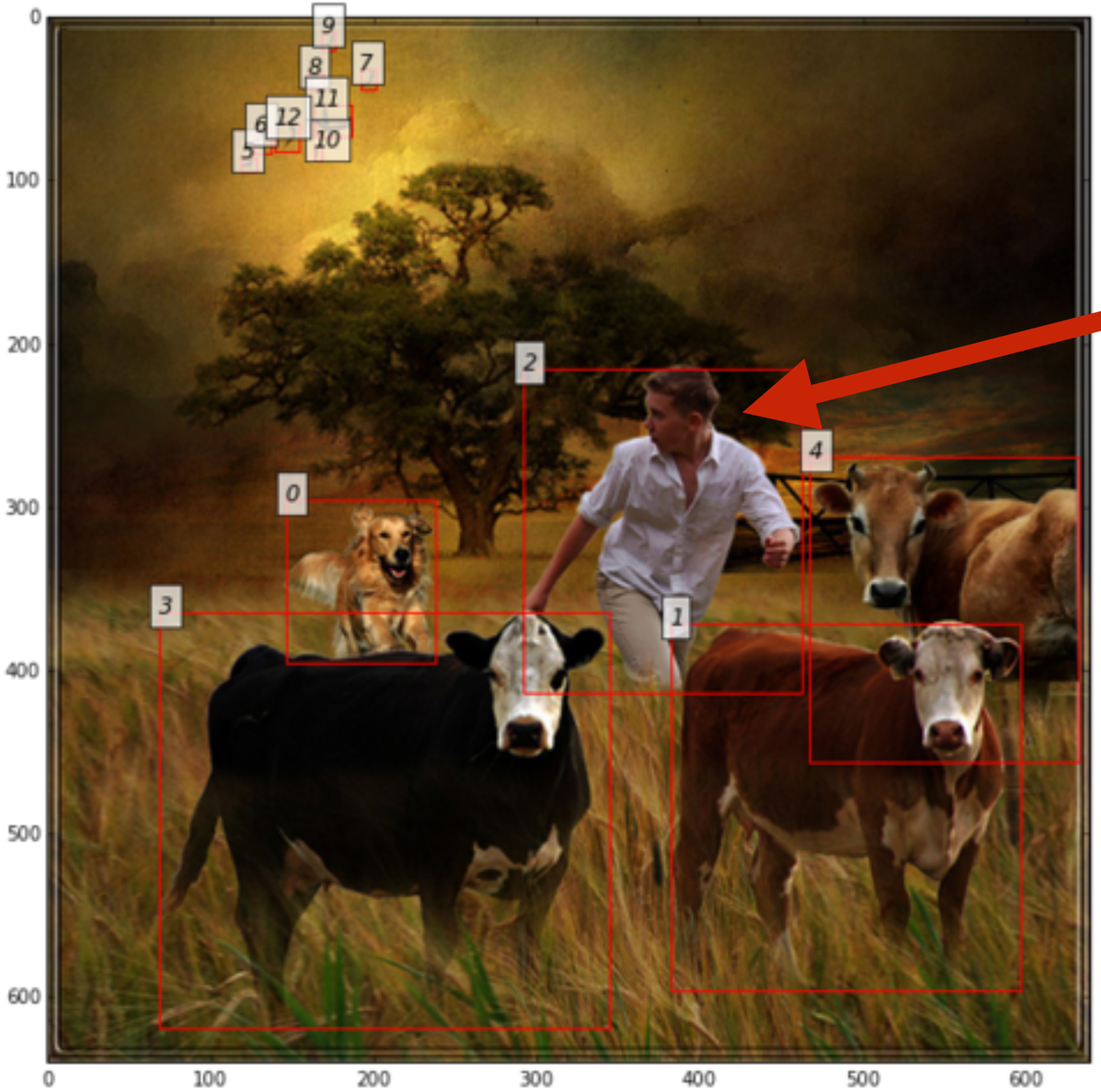
convolutional NN, trained on ImageNet 1000, fully connected layer.

In (Schlangen *et al.* 2016), GoogLeNet (Szegedy *et al.* 2015), here ResNet 50 (He *et al.* 2015)

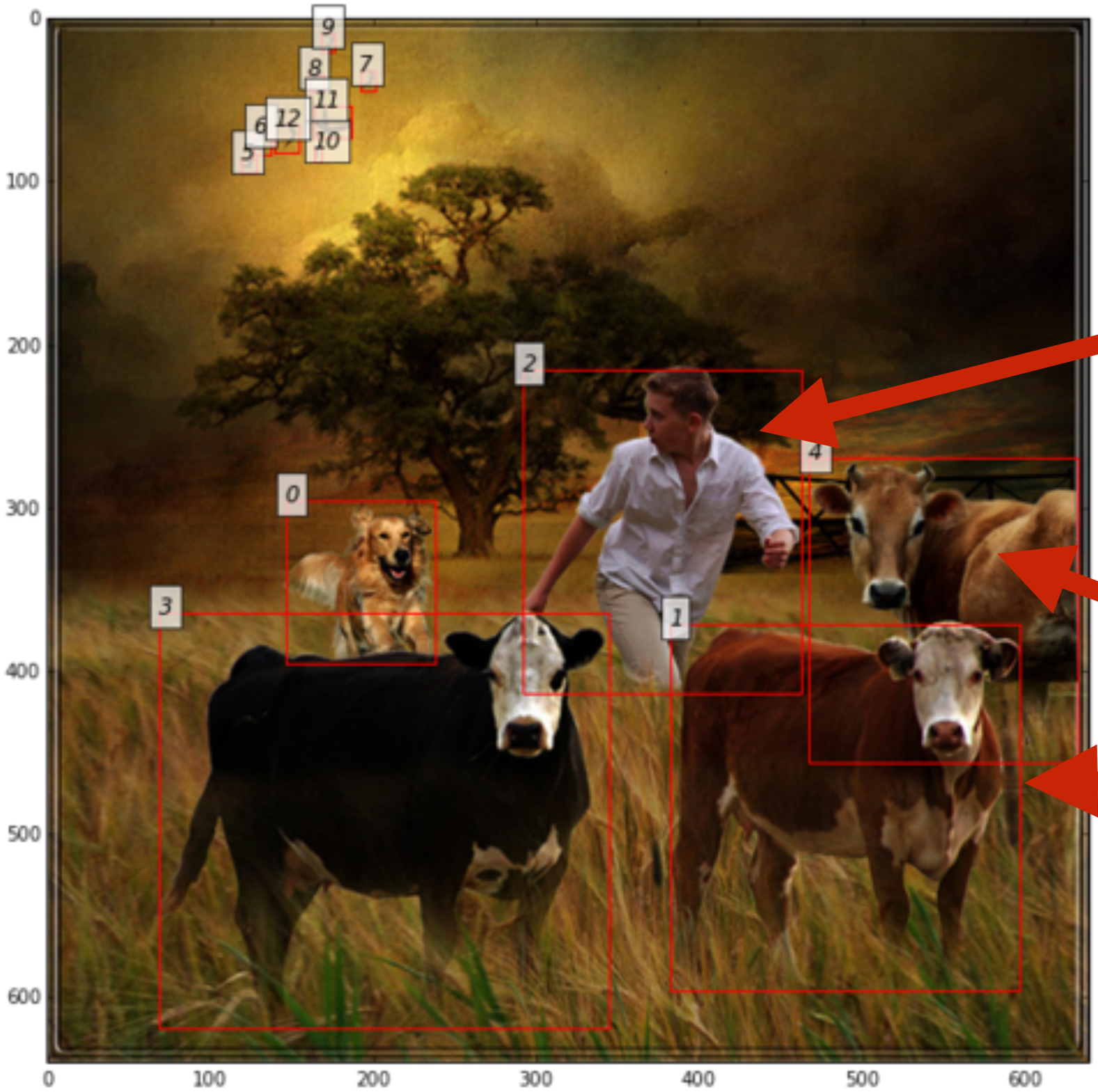
2048 CNN features
+ 7 positional features

Training

Guy with white shirt



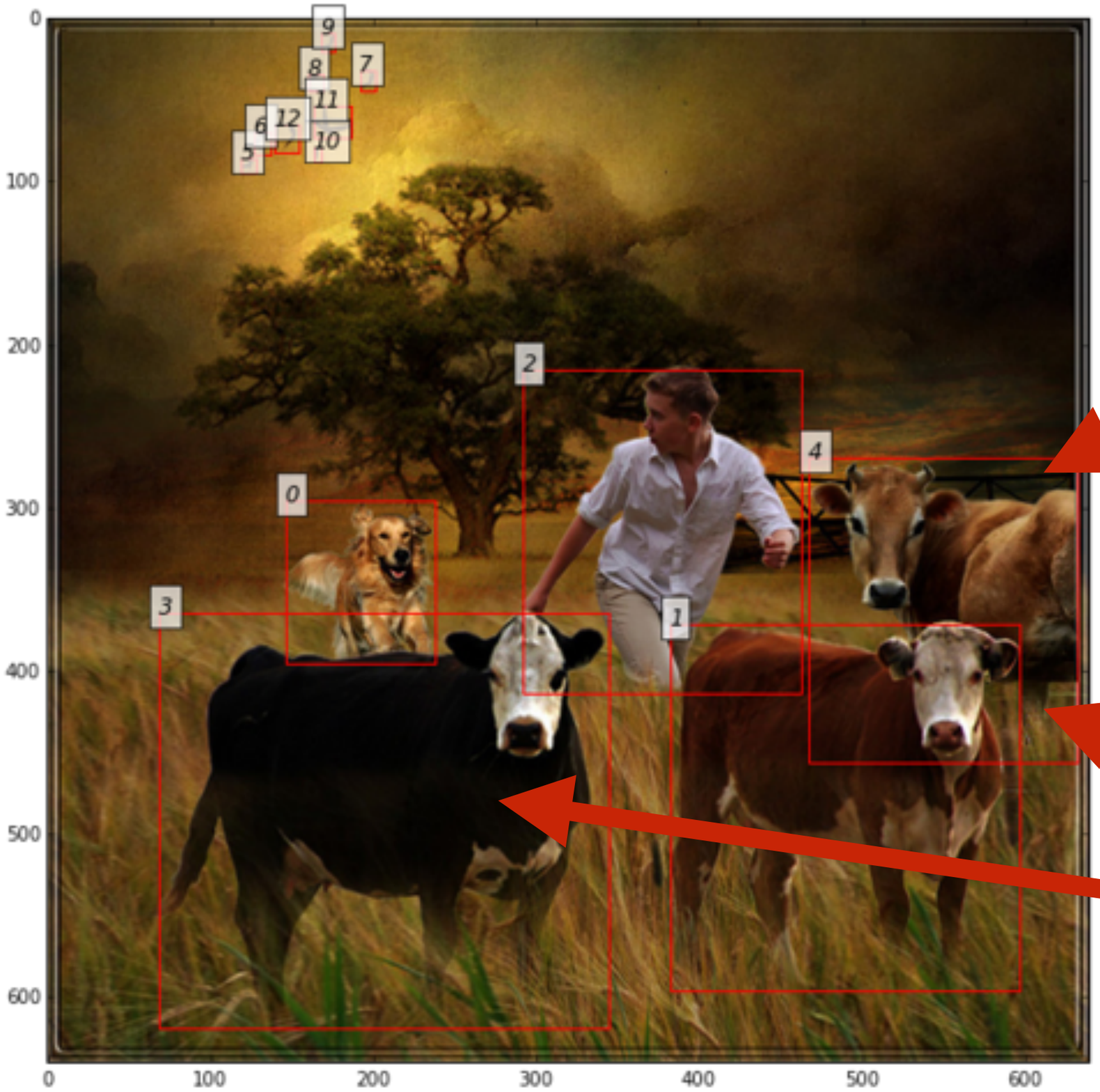
Training



Guy with white shirt

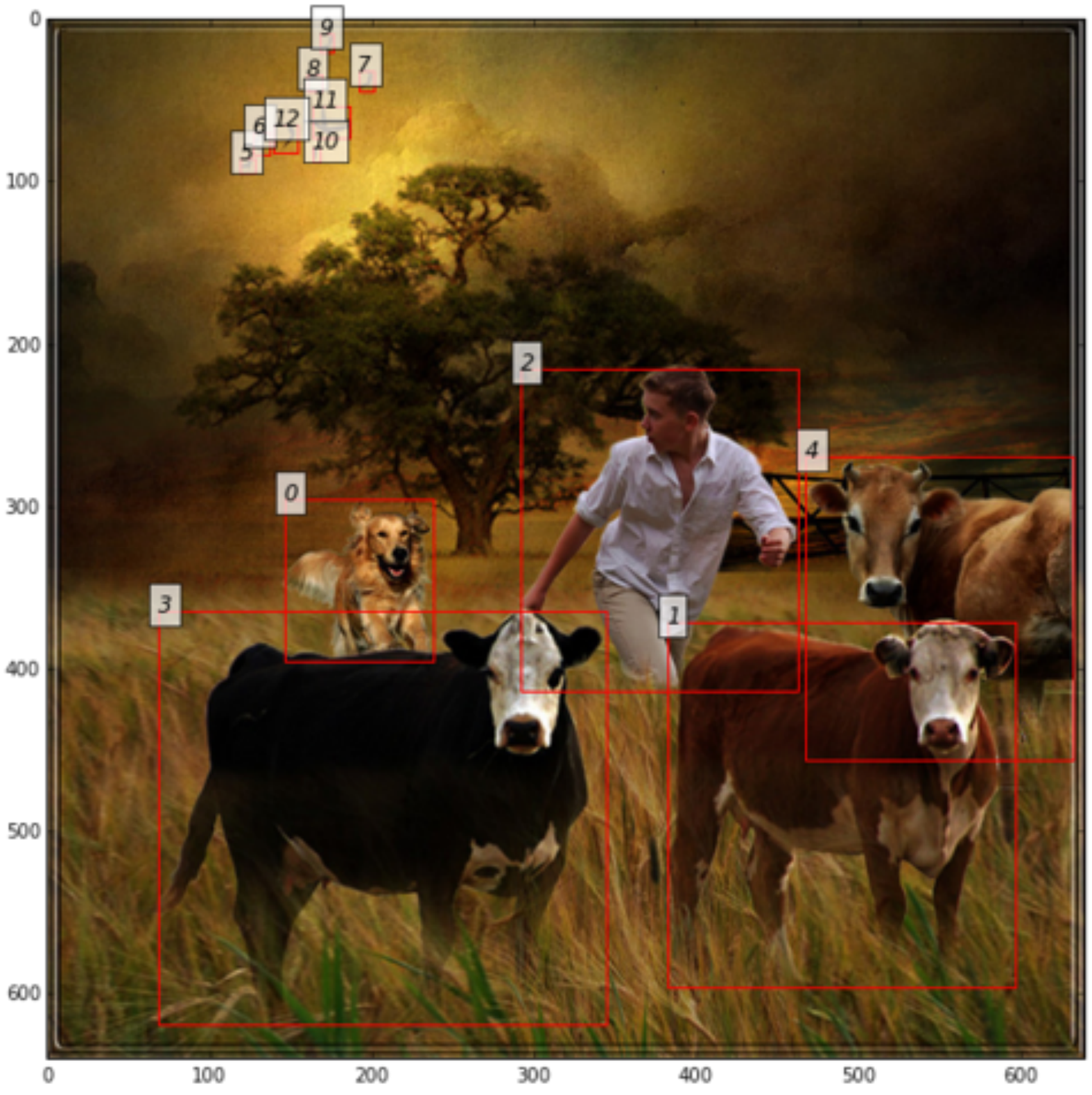
¬Guy
¬with
¬white
¬shirt

Training



Cow
right

¬Cow
¬right



Training

All words separately.

One classifier per word.

Here, trained in batch mode, but could be done incrementally.

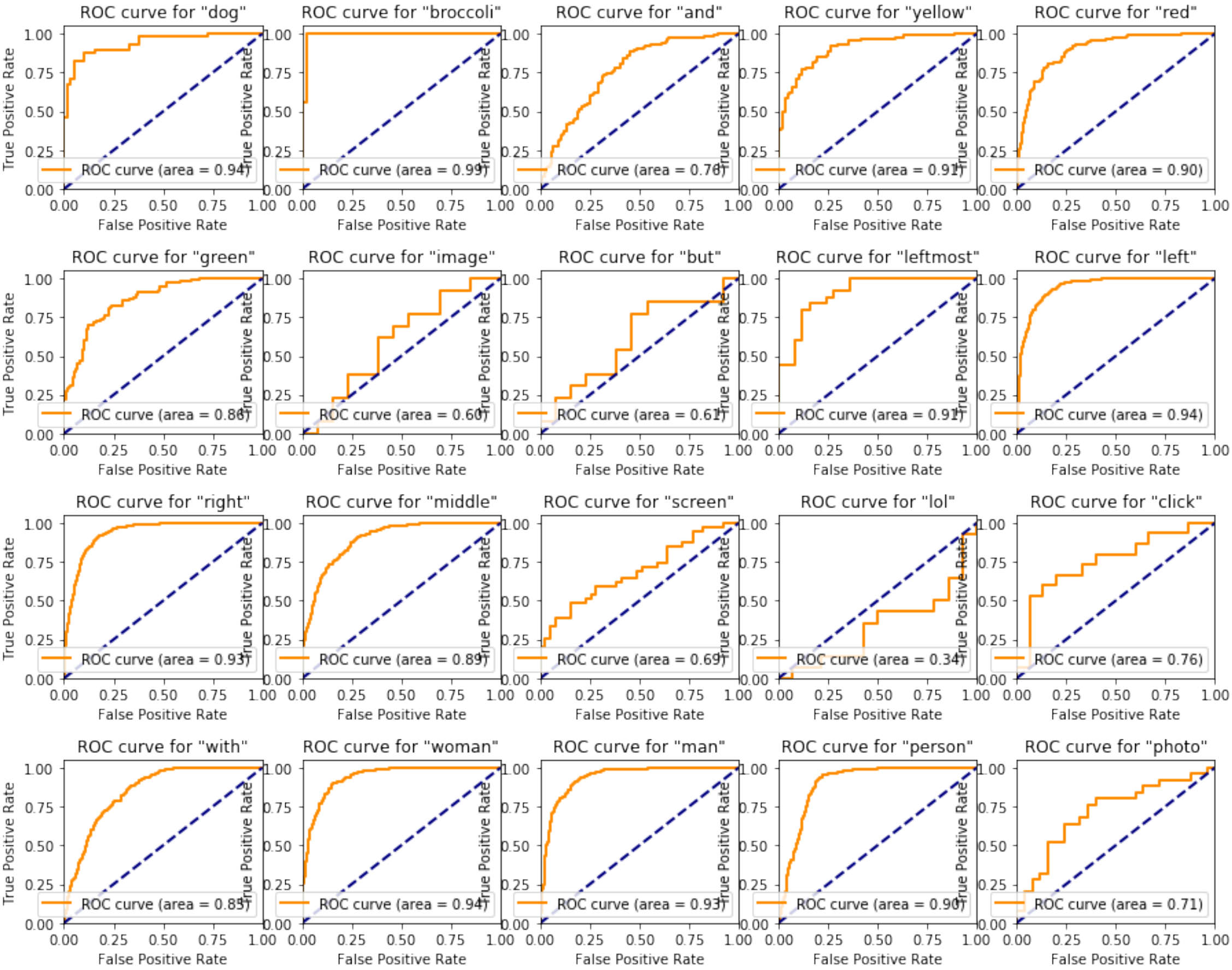
Tried taking neg inst. from same scene, and randomly from whole set.

Training

- condition: min. 40 positive training instances, splits as suggested by (Yu *et al.* 20169

- resulting vocab. size:
  - SAIAPR: 429
  - RefCoco: 503
  - SAIAPR + RefCoco: 783
  - SAIAPR + RefCoco + RefCoco$_+$: 1,174

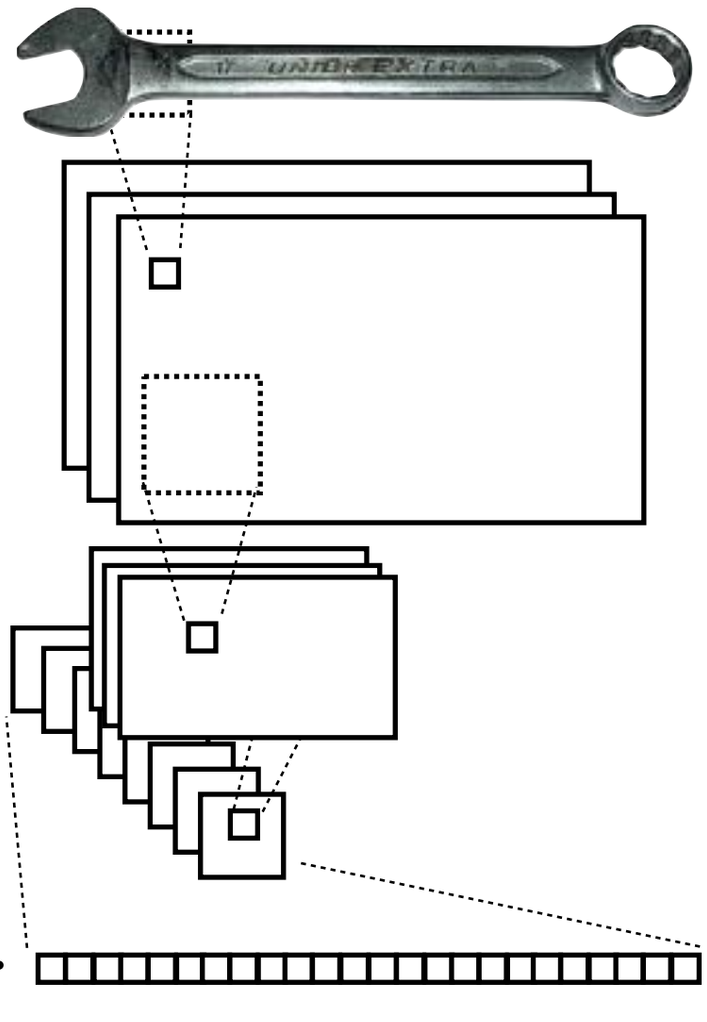
eval uation


ROC curve for "dog"
ROC curve (area = 0.94)
ROC curve for "broccoli"
ROC curve (area = 0.99)
ROC curve for "and"
ROC curve (area = 0.78)
ROC curve for "yellow"
ROC curve (area = 0.91)
ROC curve for "red"
ROC curve (area = 0.90)
ROC curve for "green"
ROC curve (area = 0.86)
ROC curve for "image"
ROC curve (area = 0.60)
ROC curve for "but"
ROC curve (area = 0.61)
ROC curve for "leftmost"
ROC curve (area = 0.91)
ROC curve for "left"
ROC curve (area = 0.94)
ROC curve for "right"
ROC curve (area = 0.93)
ROC curve for "middle"
ROC curve (area = 0.89)
ROC curve for "screen"
ROC curve (area = 0.69)
ROC curve for "lol"
ROC curve (area = 0.34)
ROC curve for "click"
ROC curve (area = 0.76)
ROC curve for "with"
ROC curve (area = 0.85)
ROC curve for "woman"
ROC curve (area = 0.94)
ROC curve for "man"
ROC curve (area = 0.93)
ROC curve for "person"
ROC curve (area = 0.90)
ROC curve for "photo"
ROC curve (area = 0.71)
True Positive Rate
False Positive Rate
0.00
0.25
0.50
0.75
1.00

application

- Now we have classifiers for the words…

- … but how do we get predictions for referring expressions from them?

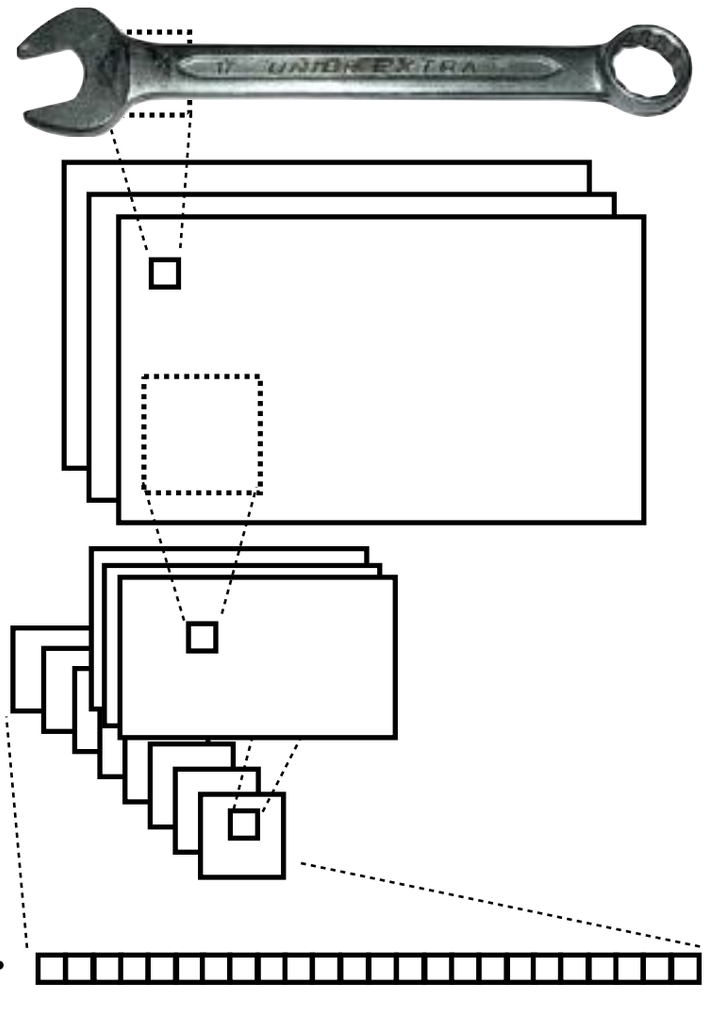


[0,1] ⊙ [0,1] ⊙ [0,1] ⊙ [0,1] ↦ [0,1]

silver wrench in middle

[0,1] ⊙ [0,1] ⊙ [0,1] ⊙ [0,1] ↦ [0,1]

[0,1] ⊙ [0,1] ⊙ [0,1] ⊙ [0,1] ↦ [0,1]

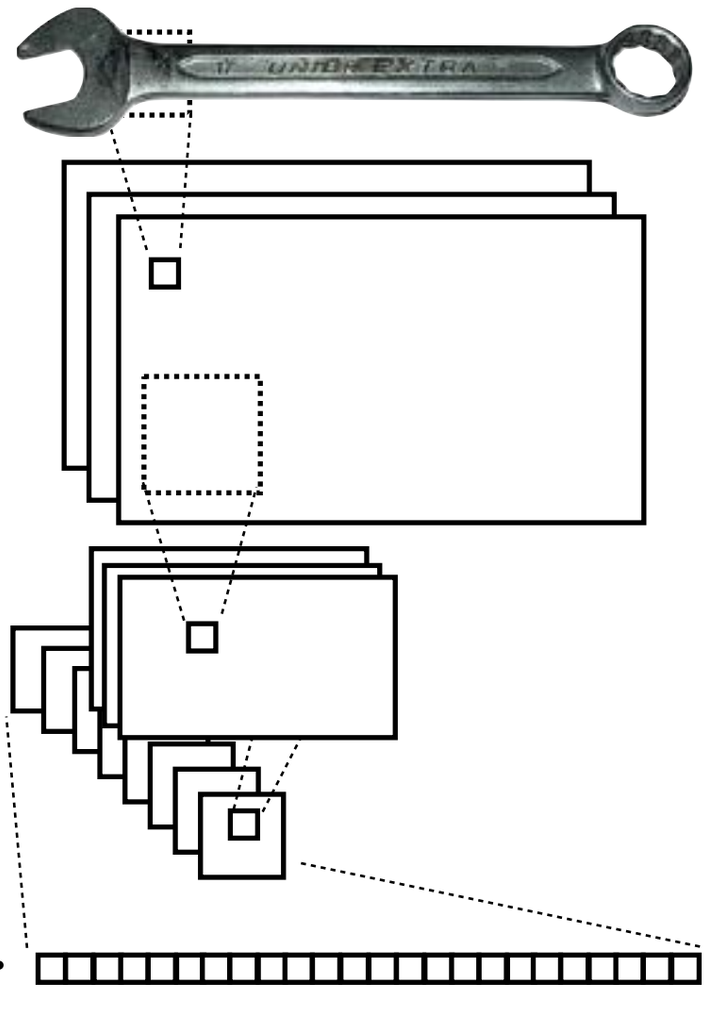


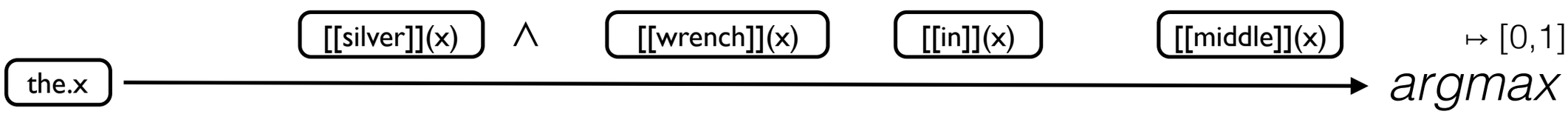
[0,1] ⊙ [0,1] ⊙ [0,1] ⊙ [0,1] ↦ [0,1]

the → silver → wrench → in → middle → *argmax*

[0,1] ⊙ [0,1] ⊙ [0,1] ⊙ [0,1] ↦ [0,1]

[0,1] ⊙ [0,1] ⊙ [0,1] ⊙ [0,1] ↦ [0,1]


σ(Θ·
T)


the.x
[[silver]](x)
∧
[[wrench]](x)
[[in]](x)
[[middle]](x)
↦ [0,1]
argmax

results

	rc testA	rc testB	rc+ testA	rc+ testB	saiapr
random	0.08	0.13	0.08	0.13	0.14
WAC @1	0.57	0.54	0.49	0.39	0.58
@3	0.87	0.87	0.84	0.81	0.85
NR, Cov, random	0.08	0.13	0.08	0.13	0.14
NR, Cov, @1	**0.60**	**0.57**	**0.51**	**0.42**	**0.63**
@3	0.90	0.89	0.86	0.84	0.88

results

			RefCOCO		RefCOCO+		RefCOCOg
		Visual Encoder	testA	testB	testA	testB	test
1	MMI (Mao et al. 2016)	VGG16	64.90	54.51	54.03	42.81	-
2	NegBag (Nagaraja, Morariu, and Davis 2016)	VGG16	58.60	56.40	-	-	49.50
3	CG (Luo and Shakhnarovich 2017)	VGG16	67.94	55.18	57.05	43.33	-
4	Attr (Liu, Wang, and Yang 2017)	VGG16	72.08	57.29	57.97	46.20	-
5	CMN (Hu et al. 2017)	VGG16	71.03	65.77	54.32	47.76	-
6	Speaker (Yu et al. 2016)	VGG16	67.64	55.16	55.81	43.43	-
7	**Spearker**+Listener+Reinforcer (Yu et al. 2017)	VGG16	72.94	62.98	58.68	47.68	-
8	Spearker+**Listener**+Reinforcer (Yu et al. 2017)	VGG16	72.88	63.43	60.43	48.74	-
9	VC(Zhang, Niu, and Chang 2018)	VGG16	73.33	67.44	58.40	53.18	-
10	ParallelAttn (Zhuang et al. 2018)	VGG16	75.31	65.52	61.34	50.86	-
11	LGRANs (Wang et al. 2019)	VGG16	76.6	66.4	64.0	53.4	-
12	DGA (Yang, Li, and Yu 2019b)	VGG16	78.42	65.53	69.07	51.99	51.99
13	**Spearker**+Listener+Reinforcer (Yu et al. 2017)	ResNet-101	73.71	64.96	60.74	48.80	59.63
14	Spearker+**Listener**+Reinforcer (Yu et al. 2017)	ResNet-101	73.10	64.85	60.04	49.56	59.21
15	MAttNet (Yu et al. 2018b)	ResNet-101	80.43	69.28	70.26	56.00	**67.01**
16	Ours	DLA-34	**81.06**	**71.85**	**70.35**	**56.32**	65.73

(Liao *et al.* 2019)

@3	0.87	0.87	0.84	0.81	0.85
NR, Cov, random	0.08	0.13	0.08	0.13	0.14
NR, Cov, @1	**0.60**	**0.57**	**0.51**	**0.42**	**0.63**
@3	0.90	0.89	0.86	0.84	0.88

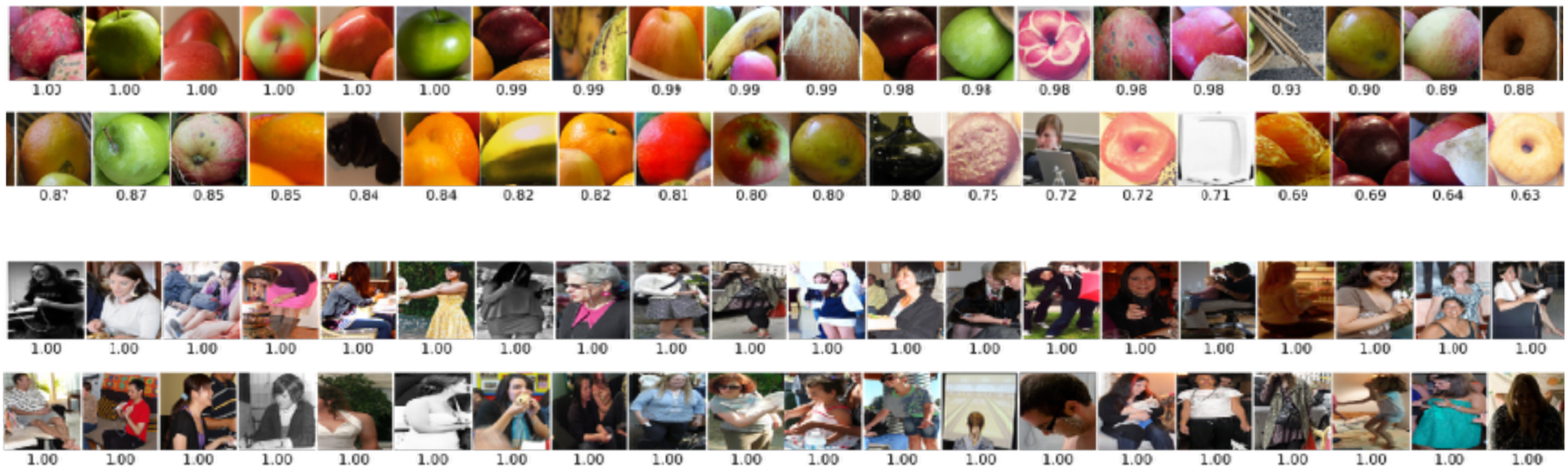
why bother?

- current state-of-the-art approaches highly fine-tuned to this task (and this dataset?), non-incremental, etc..

- we can inspect our concepts!

what do they learn?

apple



woman

yellow

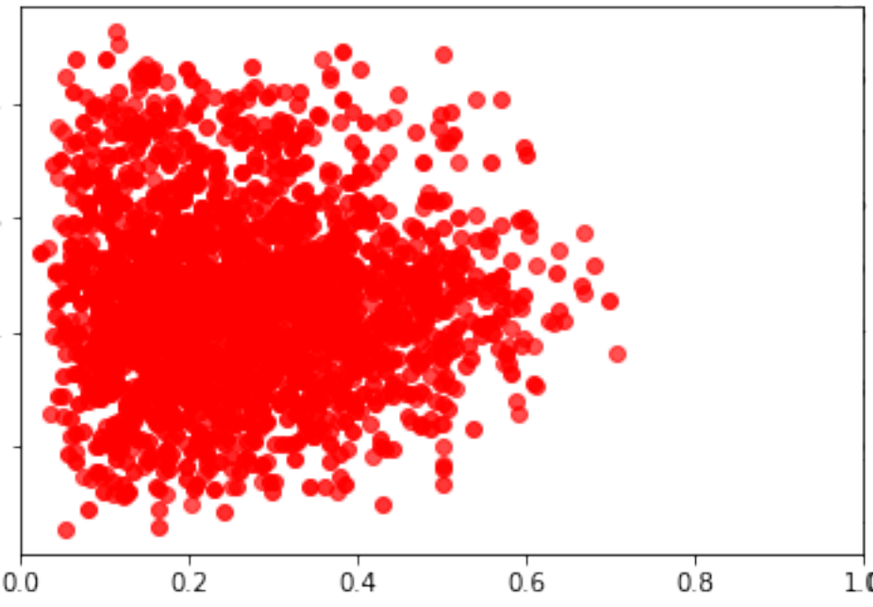
blue

what do they learn?

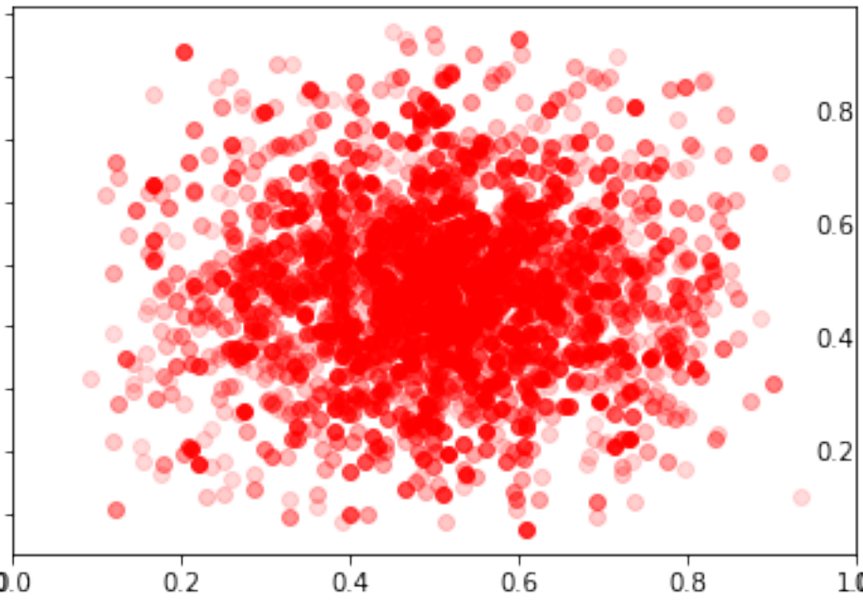
and

food

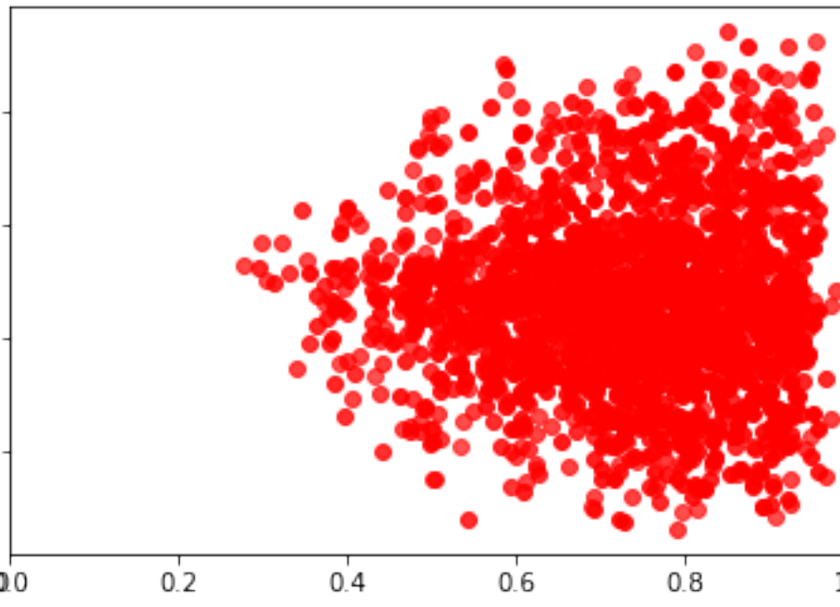
what do they learn?



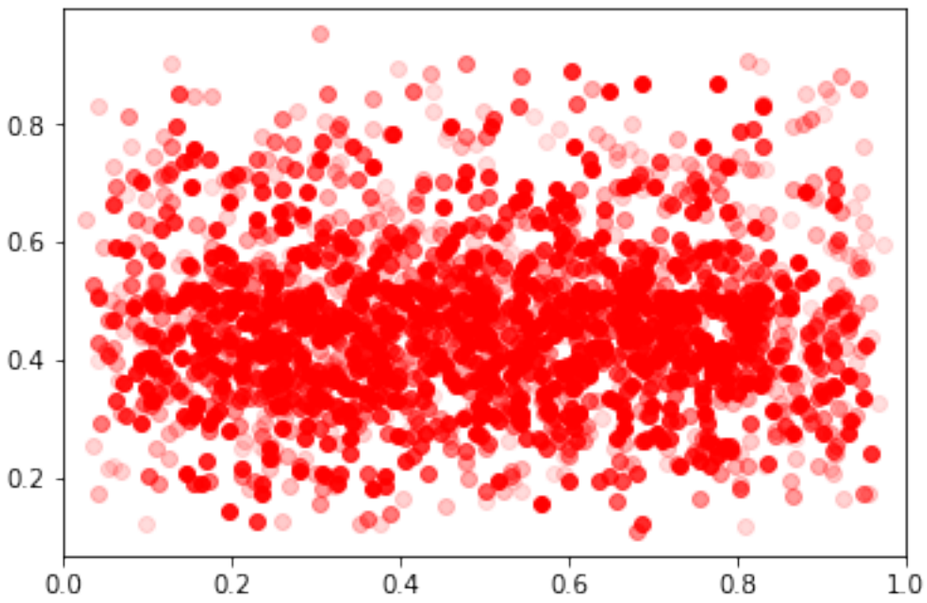
left



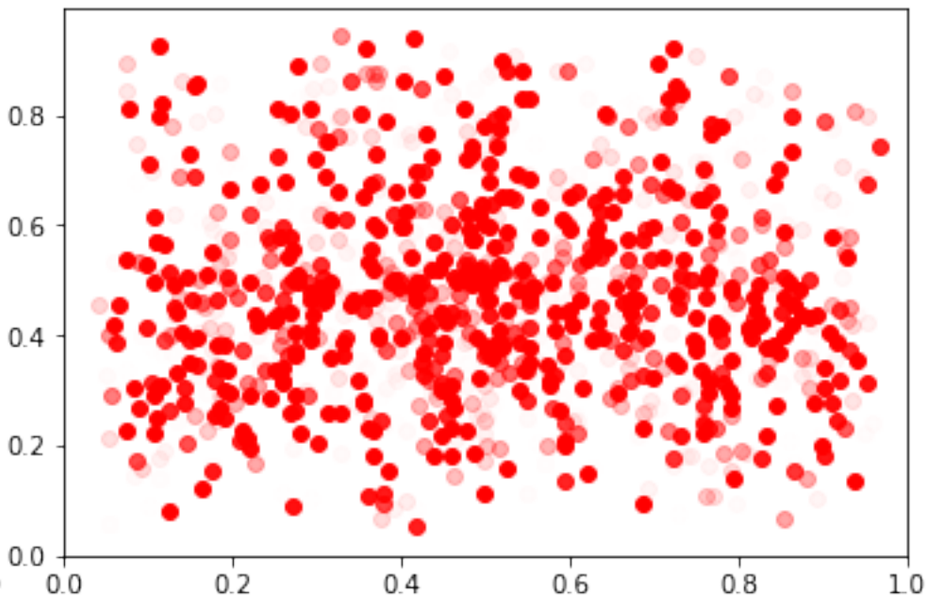
center



right



woman



yellow

what kind of model of concepts is this?

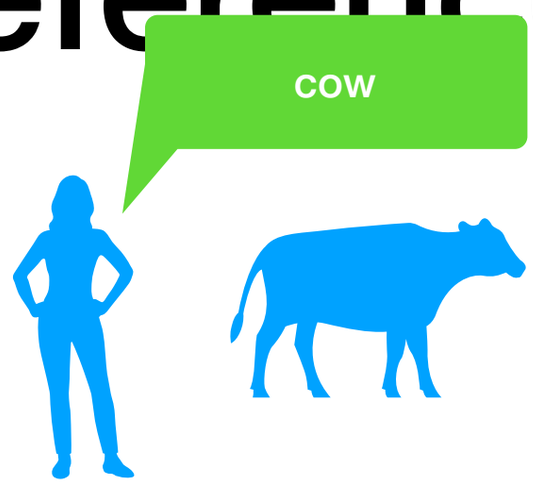
- are these things images?
    - no
- are they prototypes?
    - no, at least not if you think of prototypes as exemplars
    - they contain *negative* information as well
- are they feature bundles?
    - no, at least not if you are expecting interpretable features

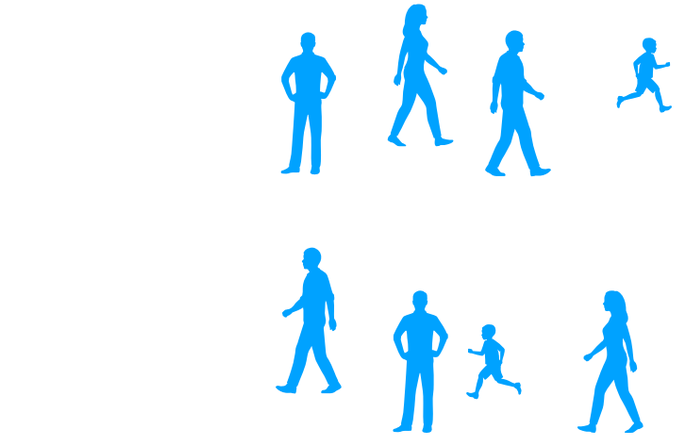
“Hey, the 1970s called, they want their fuzzy logic back!”

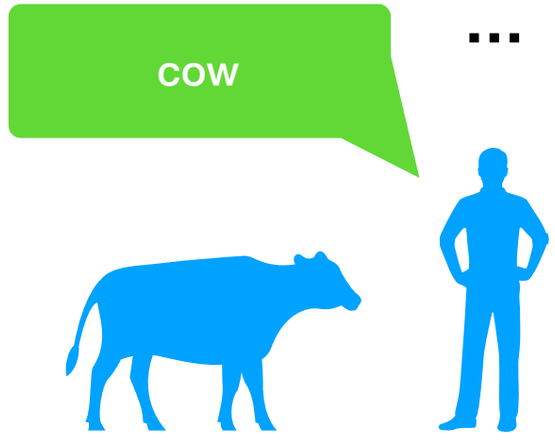
- Hang on! Real-valued set membership functions? Logical *and* as multiplication? Isn’t that just fuzzy set logic? (Zadeh 1965)
  - Well, yes, kind of. But it’s the agent’s acquired interpretation function..
- But what about *the apple that’s not an apple* (Osherson & Smith 1981; Kamp & Partee 1995)*? And does this mean that we have real-valued truth values now?
  - I don’t know yet. Maybe a pragmatic solution is possible. Confidence in interpretation, decision to believe / assert…

* $a \wedge \neg a: \varnothing$, but: $apple(o_1) : 0.1$, $\neg apple(o_1) : 0.9$

a causal-chain theory of reference?







functional *reprsnt.nal*

Conceptual Apparatus

Reference

- word to world
- categori-sation
- naming / resolution

classifiers on perceptual input

Look at the white dog!

$$[[dog]]^{D} = \{ (o, f_{dog}(o)) \}$$

what kind of function is this? (what is the range?)

where do we get this function?

how do we present image object to function?

what kind of set is this?

What about the other direction?
(Naming, Generation)

today

- data
  - what's in a language & vision corpus?
  - what's in the LV corpora that we've used?
- the "words as classifiers" model of referential concepts
  - basics: learning & simple application
  - what do they learn?
  - naming & factor graphs

functional *reprsnt.nal* *composition* *coordination*

Conceptual Apparatus

Inference

- word to word
- discourse resolution

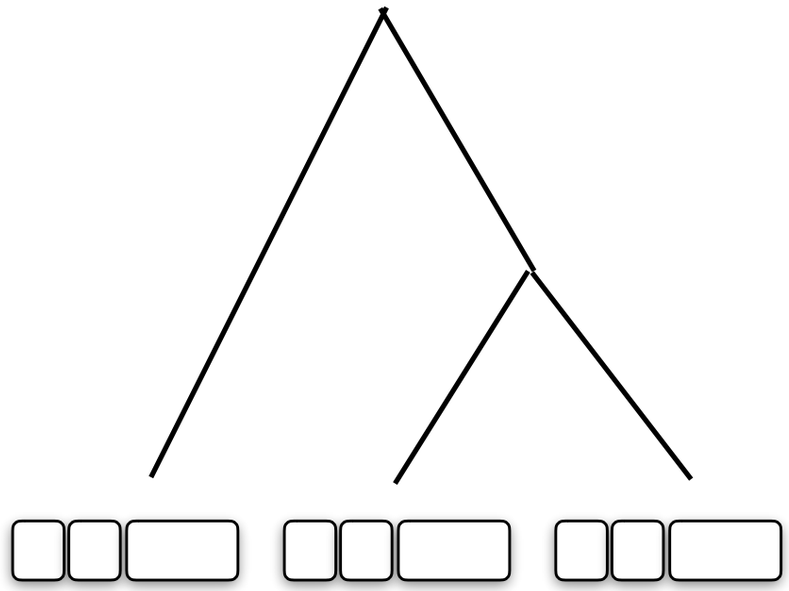
Reference

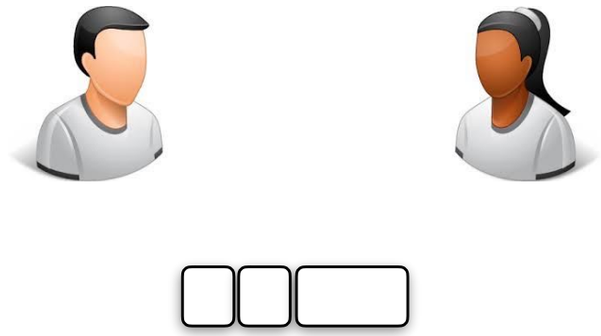
- word to world
- categori-sation
- naming / resolution

symbolic repr.

continuous repr.

classifiers on perceptual input





Look at the white dog!

We just saw a cute dog.

The cutest poodle ever!

Actually, that wasn't a poodle. It was too tall. It was a labradoodle.

- learning
  - incremental (within concept, within vocab)
  - fast
- implemented & tested on real data

Thank you.

I will be here (in Room 520) until Oct 13. Happy to meet up!

Thanks also to my Bielefeld PhD students, Postdocs, and collaborators: Julian Hough, Sina Zarrieß, Casey Kennington, Nikolai Ilinykh, Soledad Lopez, Ting Han, Nazia Attari, Spyros Kousidis.

Funding received from CITEC, DFG.

References

References to our own work can be resolved via http://clp.ling.uni-potsdam.de/publications/ (where also the PDFs are available).
(First authors: Han, Kennington, Kousidis, Lopez, Schlangen, Zarrieß.)

- Bloom, Paul (2000). How Children Learn the Meaning of Words. MIT Press
- Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollar, P., & Zitnick, C. L. (2015). Microsoft COCO Captions: Data Collection and Evaluation Server. ArXiv, 1–7.
- Forbes, M., & Choi, Y. (2017). Verb Physics: Relative Physical Knowledge of Actions and Objects. In ACL 2017.
- Glaser, W. R. (1992). Picture naming. Cognition, 42(1), 61–105.
- He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. ArXiv
- Kamp, H., & Partee, B. (1995). Prototype theory and compositionality. Cognition, 57(2), 129–191.
- Kazemzadeh, S., Ordonez, V., Matten, M., & Berg, T. L. (2014). ReferItGame : Referring to Objects in Photographs of Natural Scenes. In EMNLP 2014 (pp. 787–798).
- Kitaev, N., & Klein, D. (2018). Constituency Parsing with a Self-Attentive Encoder. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia
- Liao, Y., Liu, S., Li, G., Wang, F., Chen, Y., Qian, C., & Li, B. (2019). A Real-Time Cross-modality Correlation Filtering Method for Referring Expression Comprehension. ArXiv.
- Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A., & Murphy, K. (2016). Generation and Comprehension of Unambiguous Object Descriptions. In CVPR 2016
- Osherson, D. N., & Smith, E. E. (1981). On the adequacy of prototype theory as a theory of concepts. Cognition, 9(1), 35–58.
- Rosch, E. (1978). Principles of Categorization. In E. Rosch & B. B. Lloyd (Eds.), Cognition and Categorization (pp. 27--48). Hillsdale, N.J., USA: Lawrence Erlbaum.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2014). Going Deeper with Convolutions. arXiv preprint arXiv: 1409.4842.
- Yu, L., Poirson, P., Yang, S., Berg, A. C., & Berg, T. L. (2016). Modeling Context in Referring Expressions. In European Conference on Computer Vision (ECCV 2016).
- Zadeh, L. A. (1965). Fuzzy sets. Information and Control, 8(3), 338–353.