Vision and Feature norms: Improving automatic feature norm learning through cross-modal maps

Luana Bulat, Douwe Kiela and Stephen Clark

Computer Laboratory, University of Cambridge, United Kingdom

Feature norms

- Also referred to as property norms
- Datasets constructed by asking participants to identify the most important attributes of a concept (McRae et al., 2005; Devereaux et al., 2013)

BANANA	CELLO	SHOES
is yellow, 29	a musical instrument, 26	made of leather, 24
a fruit, 25	has strings, 16	has heels, 15
is edible, 13	made of wood, 16	worn on feet, 13
is soft, 12	found in orchestras, 13	has laces, 13
grows on trees, 11	is large, 13	worn for protection, 11
eaten by peeling, 10	requires a bow, 9	has high heels, 10

Task

- We focus on automatic prediction of conceptual properties
- Properties of concepts are important for:
  - Evaluating models of conceptual representation in cognitive science
  - Study of multi-modal semantic models (used as proxy for perceptual information)

Solution

- **Observation:** a large proportion of information contained in property norm datasets can be attributed to extra-linguistic modalities
  - >25% of features are visual (is black, is white, has a motor, made of lace)
- **Solution:** learn to predict property norms through cross-modal mappings from raw perceptual information (i.e. image data)

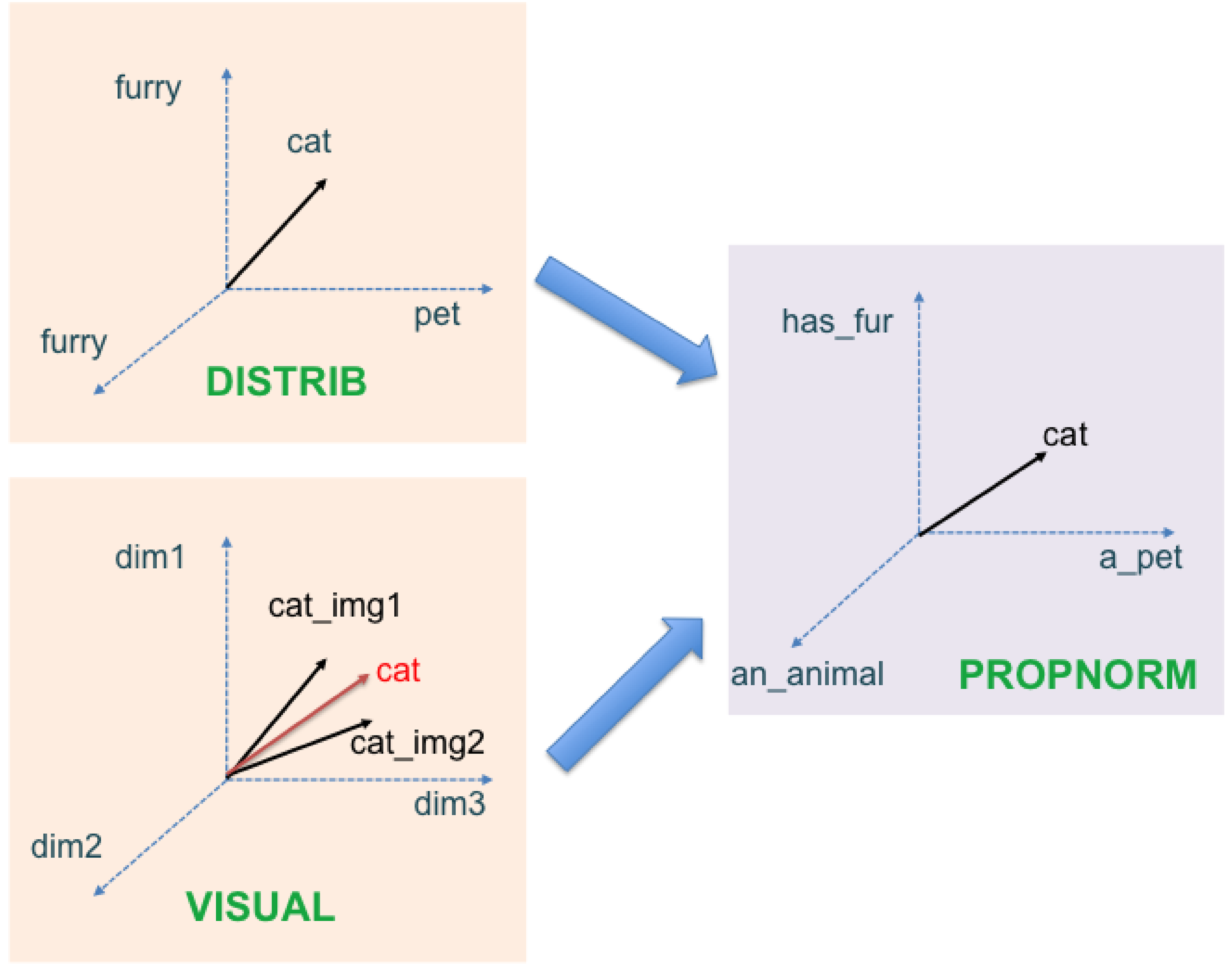
Data

For all McRae concepts we obtain:

- Attribute-based representations
  - PROPNORM: by treating the McRae dataset as a bag-of-properties (dimensions: properties, counts: production frequencies)
- Image representations
  - VISUAL: by extracting the pre-softmax layer from a forward pass in a convolutional neural network that has been trained on the ImageNet classification task using Caffe (Jia et al., 2014)
- Linguistic representations
  - DISTRIB: 10K dimensional count-based distributional vectors
  - SVD: 300-dimensional SVD reduced version of DISTRIB
  - EMBED: pre-trained word embeddings (Mikolov et al., 2013)
- Multi-modal representations (concatenation of L2-normalised visual and linguistic vectors):
  - VISUAL+DISTRIB, VISUAL+SVD, VISUAL+EMBED

Method

- We use partial least squares regression (PLSR) to learn cross-modal maps to PROPNORM from the visual (VISUAL), linguistic (DISTRIB, SVD, EMBED) and multi-modal spaces (VISUAL+DISTRIB, VISUAL+SVD, VISUAL+EMBED)



- We use the zero-shot learning procedure: for each of the 541 concepts in McRae, we train a mapping on the remaining 540 concepts and record whether the correct label is retrieved among the top N neighbours

Evaluation

- Average percentage correct at N: proportion of test instances that are ranked within the top N highest ranked nearest neighbors.

Experimental Results

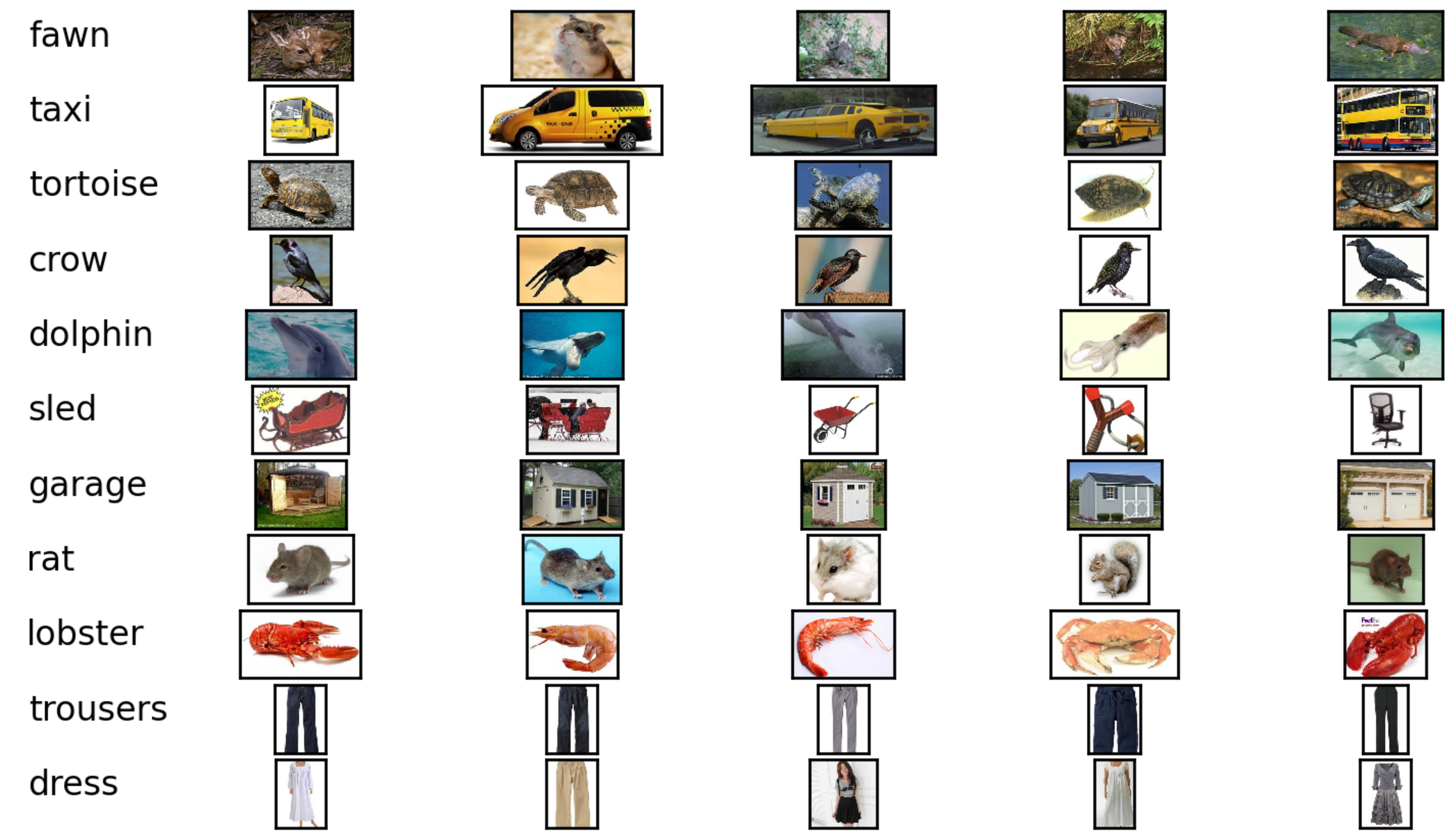
From	P@1	P@5	P@10	P@20
DISTRIB	1.30	6.88	16.54	26.58
SVD	2.79	22.12	38.10	**57.99**
EMBED	3.90	23.42	36.80	55.02
VISUAL	3.35	28.44	47.96	**64.50**
VISUAL+DISTRIB	2.60	23.23	39.41	56.13
VISUAL+SVD	2.97	28.44	50.74	**65.43**
VISUAL+EMBED	3.16	28.44	51.12	65.06
RANDOM	0.0	0.74	2.42	3.90

Qualitative results

Concept	Nearest neighbours (from VISION)
banana	**banana**, lemon, corn, pear, grapefruit, pineapple
crocodile	alligator, **crocodile**, frog, turtle, iguana, toad
cello	violin, guitar, banjo, harp, harpsichord, **cello**, flute
drum	pot, pan, coin, skillet, bucket, peg, cap (bottle)
fox	**fox**, cougar, coyote, deer, mink, elk, chipmunk
harpoon	sword, machete, **harpoon**, dagger, rifle, knife, gun
pants	jeans, trousers, **pants**, shirt, blouse, jacket, coat
Concept	**Top predicted features (from VISION)**
banana	is yellow, is black*, is round, is long, a fruit
crocodile	is green, an animal, lives in water, beh - swims
cello	has strings, a musical instrument, made of wood
drum	made of metal, is round, used for cooking*
fox	an animal, is fast, is small, has fur, has a tail
harpoon	made of metal, a weapon, is sharp, is dangerous*
pants	clothing, has buttons, is blue*, different colours

Property based query engine

- We can learn how concepts look like based on their properties, by learning a cross-modal map from PROPNORM to VISUAL.



Conclusion

- Raw visual information (images) is a better predictor for conceptual properties than linguistic input (text).

UNIVERSITY OF CAMBRIDGE

EPSRC

Computer Laboratory, University of Cambridge

e-mail: ltf24@cam.ac.uk

www: http://www.cl.cam.ac.uk/~ltf24