An introduction to Markov logic networks and their use in visual relational learning

Willie Brink

Applied Mathematics, Stellenbosch University

wbrink@sun.ac.za

Thanks to Luc De Raedt and the DTAI research group at KU Leuven

Elephants are large grey animals with big ears.

Visual queries

`I see something large and grey with big ears; what is it?`

→ object recognition from **visual attributes**

`What do animals look like?`

→ visual attribute prediction from **categorical attributes**

`I see a round and red object being eaten; what is it?`

→ object recognition from **visual attributes** and **affordances**

`I have not seen this object before; what can I do with it?`

→ (zero-shot) affordance prediction from visual attributes

Attributes and affordances

Visual attributes: mid-level semantic visual concepts shared across classes[1], e.g. `furry`, `striped`, `has_eyes`, `young`

Physical attributes: e.g. `size`, `mass`, `odor`

Categorical attributes: hierarchies of semantic generalizations, e.g. `cat`, `mammal`, `animal`

Relative attributes[2]

Object affordances: possible actions that can be applied to the object[3], e.g. `grasp`, `lift`, `sit_on`, `feed`, `eat`

[1] Feris, Lampert, Parikh, *Visual Attributes*, Springer, 2017.

[2] Kovashka, Parikh, Grauman, *WhittleSearch: image search with relative attribute feedback*, CVPR, 2012.

[3] Zhu, Fathi, Fei-Fei, *Reasoning about object affordances in a knowledge base representation*, ECCV, 2014.

Relations

Relations (pos. or neg.) between attributes and affordances can lead to an expressive and semantically rich description of our knowledge, and facilitate visual reasoning.

attribute-attribute	e.g. an object with a tail likely also has a head
attribute-affordance	e.g. a spiky object is perhaps not touchable
affordance-affordance	e.g. an edible object is probably also liftable

Relations should be **statistical** and **learnable**[4].

[4] De Raedt, Kersting, *Statistical Relational Learning*, Springer, 2011.

A unified framework

We want to model these types of relations, learn about them from data, and perform inference tasks.

Separate classifiers to label objects, recognize attributes and affordances, etc.

Instead, let's consider a unified **knowledge graph** approach that

1. models the relations between attributes and affordances, and
2. enables a diverse set of visual inference tasks.

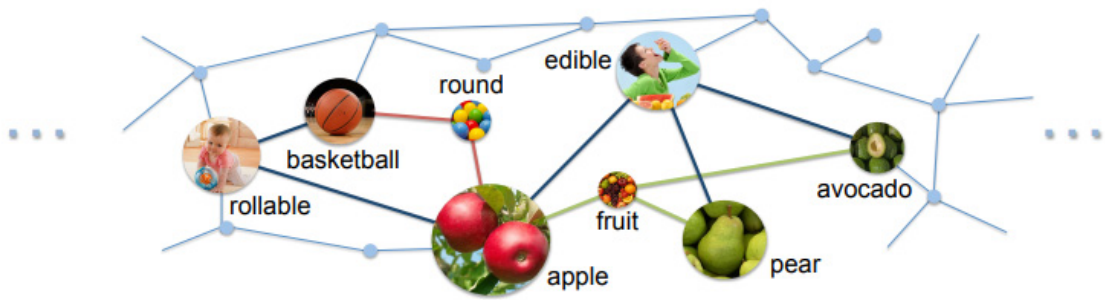


image credit: Zhu et al. (2014)

Probabilistic logic

First-order logic: convenient for expressing and reasoning about relations

e.g. apples are fruit, fruit are edible, ∴ apples are edible.

But logic is brittle.

Probabilistic models: offer a principled way of dealing with uncertainty

e.g. apples are fruit, *some* fruit are edible, ∴ this apple *might* be edible.

Markov logic networks: apply probabilistic learning and inference to the full expressiveness of first-order logic[5].

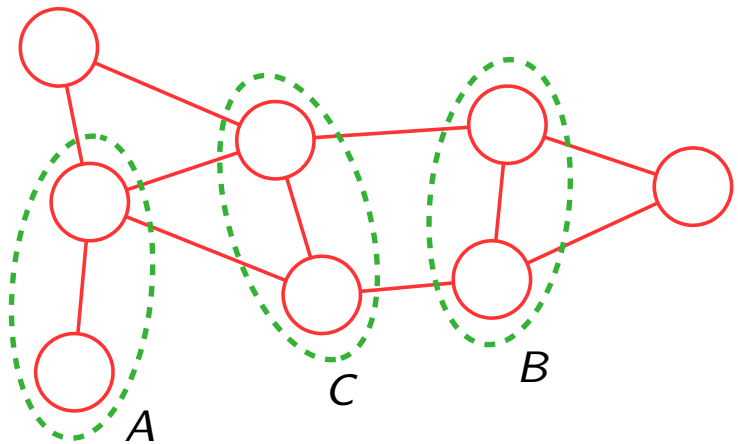
MLNs are robust, reusable, scalable, cost-effective, and human-friendly, and possess a rich relational template structure.

[5] Richardson, Domingos, *Markov logic networks*, Machine Learning, 2006.

Markov networks

. . . also called **Markov random fields** or **undirected graphical models**.

Set of random variables (nodes) and pairwise connections (edges).
Satisfies the Markov conditional independence properties.



Joint distribution factorizes over the cliques:

$$P(\mathbf{x}) = \frac{1}{Z} \prod_C \phi_C(\mathbf{x}_C), \quad \text{with} \quad Z = \sum_{\mathbf{x}} \prod_C \phi_C(\mathbf{x}_C)$$

Markov networks

Canonical exponential form:

$$\text{define} \quad E(\mathbf{x}_C) = -\log \phi(\mathbf{x}_C), \quad \text{then} \quad P(\mathbf{x}) = \frac{1}{Z} \exp\left(-\sum_C E_C(\mathbf{x}_C)\right)$$

Inference over a Markov net:

e.g. to compute the marginal of a set of variables, given values of another

exact: sum over all possible assignments to the remaining variables

approximate: loopy belief propagation, MCMC, variational Bayes, ...

First-order logic

Variable	`X`
Constant	`john`
Functor	`mother_of(X)`
Atom	`person(X)`, `friends(X,Y)`
Clause	`friends(X,Y) => [smokes(X) <=> smokes(Y)]`
Theory	set of clauses that implicitly form a conjunction
Grounded theory	contains no variables
Possible world	assignment of values to all atoms in a grounded theory

We can think of clauses with variables as **templates**.

Markov logic networks

An MLN is a set of **weighted logical clauses**.

The weight w_i specifies the strength of clause i.

MLNs can encode contradicting clauses.

If an assignment of values does not satisfy a clause, it becomes less probable, but not necessarily impossible.

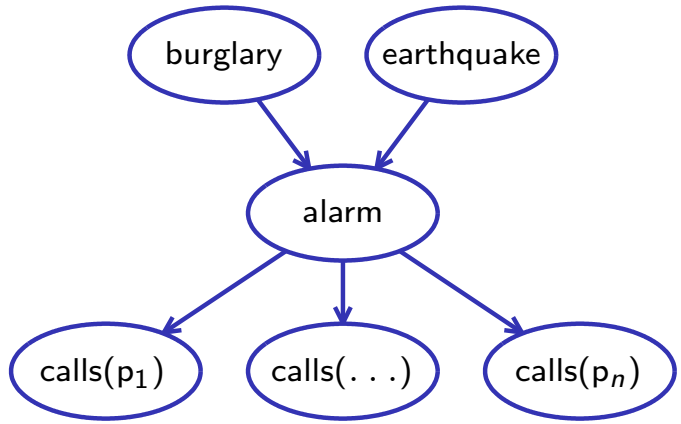
Clauses with variables are **templates** for a Markov network.

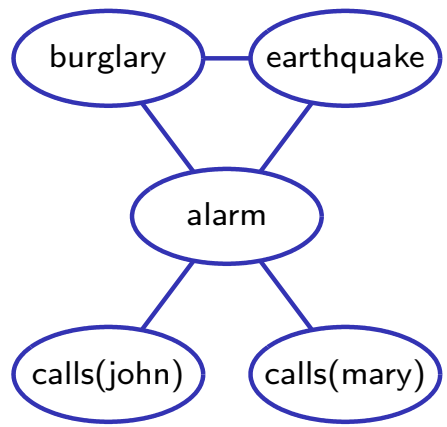
By assigning constants to all variables, we induce a **grounded Markov net**, which defines a distribution over the possible worlds.

Markov logic networks

The famous earthquake example[6]:

```
0.7 burglary
0.2 earthquake
0.9 alarm <= burglary ∧ earthquake
0.8 alarm <= burglary ∧ ¬earthquake
0.1 alarm <= ¬burglary ∧ earthquake
0.8 calls(X) <= alarm ∧ person(X)
0.1 calls(X) <= ¬alarm ∧ person(X)
1.0 person(john)
1.0 person(mary)
evidence(calls(john),true)
evidence(calls(mary),true)
query(burglary)
```





[6] Pearl, *Probabilistic Reasoning in Intelligent Systems*, Morgan Kauffman, 1988.

Inference over an MLN

Knowledge based model construction

1. ground the MLN: bipartite MN with (grounded) atoms and clauses
2. belief propagation: pass messages between atoms and clauses

This does not scale particularly well...

Lifted inference

MLNs have templates: compact representation of types of relations.

- we cluster atom-clause pairs that would pass the same messages
- only pass messages between clusters

If appropriately scaled, this is equivalent to message passing in the full grounded MN[7].

[7] Singla, Domingos, *Lifted first-order belief propagation*, AAAI Conf. on AI, 2008.

Learning in an MLN

We might want to **learn the weights** in an MLN from data.

(It is also possible to learn the structure[8].)

Closed-world assumption: what is not known to be true, is false.

Maximum likelihood estimation (similar for MAP)

Gradient ascent; turns out that $\frac{\partial}{\partial w_i} \log(P(\mathbf{y}|\mathbf{x})) = n_i(\mathbf{x}) - E_{\mathbf{y}}[n_i(\mathbf{y})]$

$n_i(\mathbf{x})$: number of times clause i is true in the data

$E_{\mathbf{y}}[n_i(\mathbf{y})]$: expected number of times clause i is true according to the model

Inference is required at every step, to calculate gradients.

[8] Kok, Domingos, *Learning the structure of Markov logic networks*, ICML, 2005.

Case study: Zhu et al. (2014)

Evidence collection

40 object and 14 affordances from the Stanford 40 Actions dataset

sample 100 images per object from ImageNet

33 pre-trained visual attribute classifiers[9]

[9] Farhadi, Endres, Hoiem, Forsyth, *Describing objects by their attributes*, CVPR, 2009.

Case study: Zhu et al. (2014)

Evidence collection

40 object and 14 affordances from the Stanford 40 Actions dataset

sample 100 images per object from ImageNet

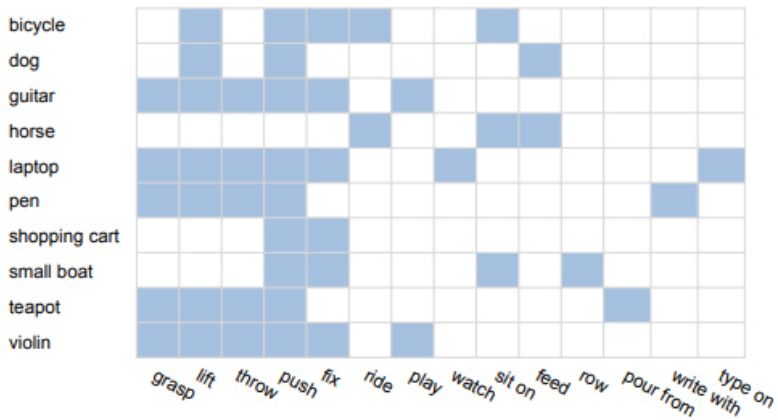
33 pre-trained visual attribute classifiers

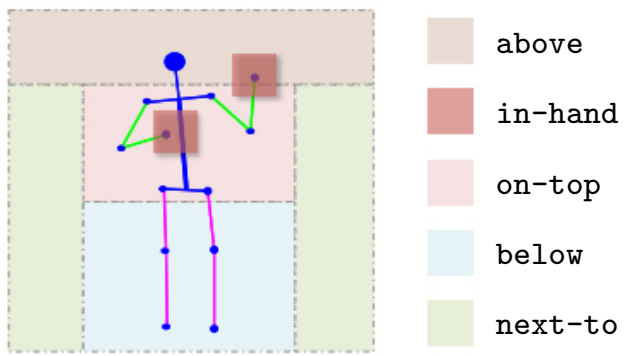
extract object weights and sizes from product details on Amazon

extract hypernym hierarchies from WordNet for categorical attributes

manually link objects with affordance labels

also describe affordance by human pose and object location





Case study: Zhu et al. (2014)

Evidence collection

40 object and 14 affordances from the Stanford 40 Actions dataset

sample 100 images per object from ImageNet

33 pre-trained visual attribute classifiers

extract object weights and sizes from product details on Amazon

extract hypernym hierarchies from WordNet for categorical attributes

manually link objects with affordance labels

also describe affordance by human pose and object location

Learning a knowledge base

define template clauses between the various types of variables

learn weights from the evidence

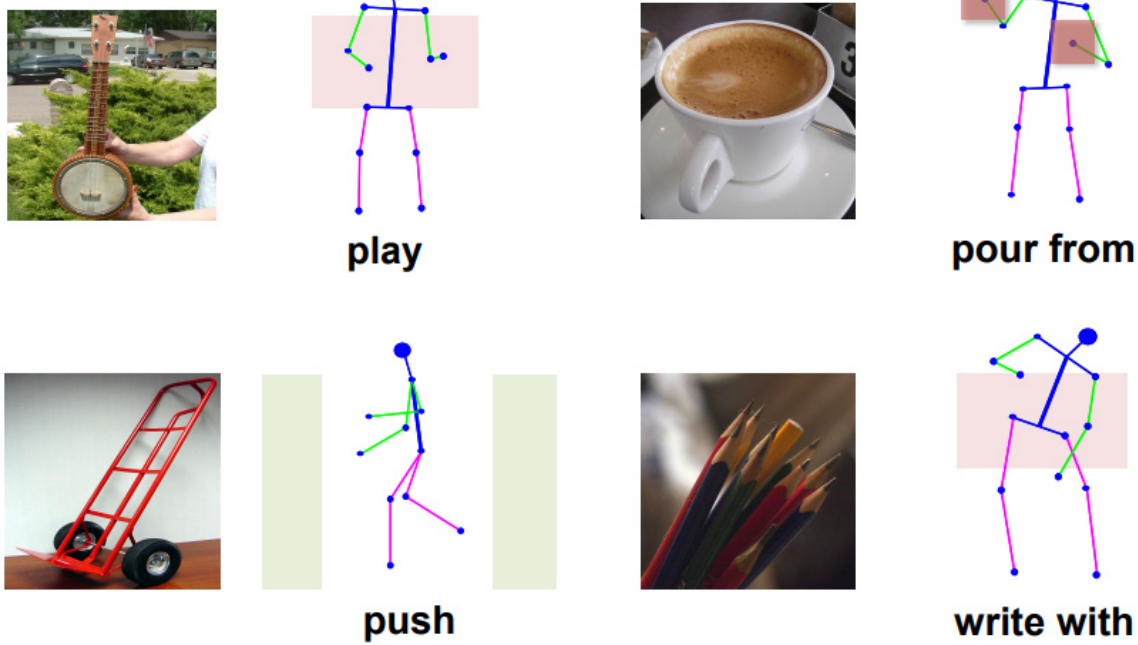
Case study: Zhu et al. (2014)

Zero-shot affordance prediction

image of a novel object

extract visual attributes and infer physical and categorical attributes

query MLN for most likely affordance, human pose and object location



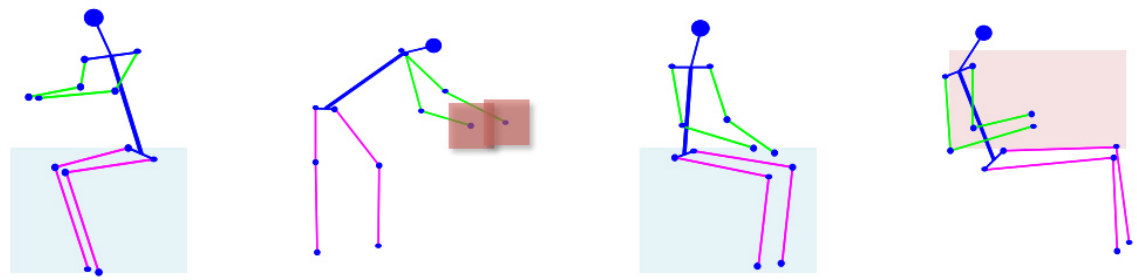
Case study: Zhu et al. (2014)

Predictions from human interaction

image of a person interacting with an object

extract human pose and object location as evidence

query MLN for most likely affordance and state of each object attribute, and retrieve object label from attributes



Further reading

- large-scale, multimodal (vision & text) knowledge base[10]
- never-ending image learning from the web[11]
- visual question answering
- discovering visual attributes in deep convolutional neural nets[12]

[10] Zhu, Zhang, Ré, Fei-Fei, *Building a large-scale multimodal KB system for answering visual queries*, CVPR, 2015.

[11] Chen, Shrivastava, Gupta, *NEIL: extracting visual knowledge from web data*, ICCV, 2013.

[12] Shankar, Garg, Cipolla, *Deep-carving: discovering visual attributes by carving deep neural nets*, CVPR, 2015.