Visually-Grounded Bayesian Word Learning

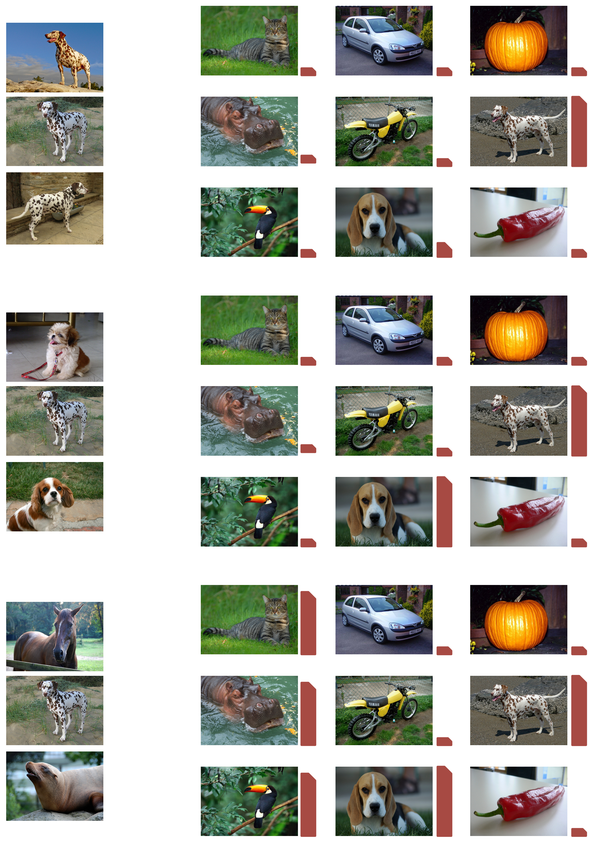
Yangqing Jia, Joshua Abbott, Joseph Austerweil, Thomas Griffiths, Trevor Darrell
Dept of EECS & Dept of Psychology, UC Berkeley

1. Motivation

- Learning Novel Concepts
  - Learning a language is one of the classic problems that is solved better by the human mind than by any computer.
  - Human learn the meanings of words accurately from just a handful of labelled examples.
  - It is a significant challenge to learn novel nouns - one simple aspect of the problem above.
- Combining Cognitive Science with Vision
  - Bayesian word learning answers the challenge using Bayesian inference to identify the intended referent of a novel noun.
  - Such cogscience models do not have a perceptual component and, instead, assume a fixed set of perfectly-recognized stimuli.
  - We show that integrating Bayesian word learning with computer vision leads to a system capable of approximating how people learn nouns directly from images.

2. Word learning

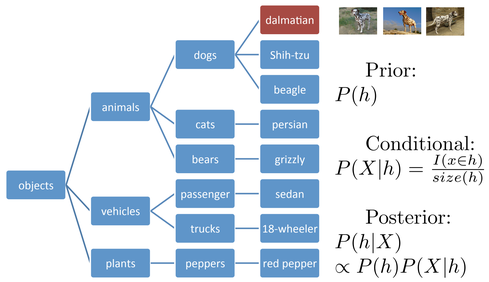
Learning a new word "DAK" from a few examples:



3. The Bayesian Word Learning Model

- We assume a set of examples of a novel concept defined by a novel word (like "fep").
- We assume a set $\mathcal{H}$ of hidden hypotheses - concept candidates.
- The goal is to predict if a new example belongs to the concept or not.
- The probability could be computed as

$$P(x_{\text{new}} \in \mathcal{C}|\mathcal{X}) = \sum_{k=1}^{K} P(x_{\text{new}} \in \mathcal{C}|h_k)P(h_k|\mathcal{X})$$



Prior:
$P(h)$

Conditional:
$P(X|h) = \frac{I(x \in h)}{size(h)}$

Posterior:
$P(h|X)$
$\propto P(h)P(X|h)$

4. Visual Grounding

- For an arbitrary image, we do not know the basic concept (leaf node) associated to it.
- Solution: combining the hypothesis hierarchy with the computer vision classifier.

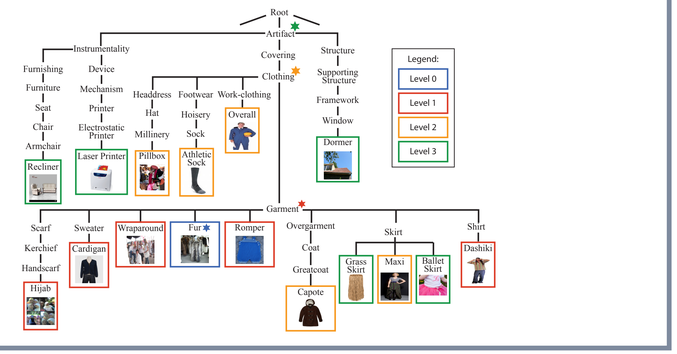
- We adopt the state-of-the-art image classification pipeline, and produce probabilistic outputs.
- The conditional is then computed as

$$P(I_n|h_k) = E_{P(x_n|I_n)}\left[P(x_n|h_k)\right] = \sum_{x=1}^{L} A_{x,\hat{x}_n} P(x_n|h_k)$$

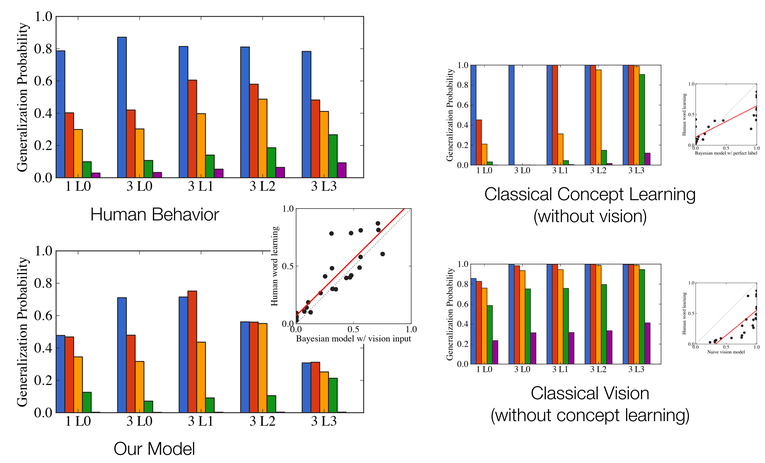
$$P(h_k|\mathcal{I}) \propto P(h_k) \prod_{n=1}^{N} P(I_n|h_k)$$

5. Experiments

- We adopted the Imagenet hierarchy to automatically generate domains of hierarchical concepts.
- Five types of examples:
  1. One example from the most specific (Level 0) category.
  2. Three examples from the most specific (Level 0) category.
  3. Three examples from categories 10% of the path to the root (L1).
  4. Three examples from categories 25% of the path to the root (L2).
  5. Three examples from categories 50% of the path to the root (L3).



6. Results



7. Remarks

Possible Future Work:

- Learn unknown hypotheses / hierarchy from human behavior.
- Learn perceptual similarity (like distance metric learning).
- Learn attributes from human behavior.

- J.T. Abbott, J.L. Austerweil, and T.L. Griffiths. Constructing a hypothesis space from the Web for large-scale Bayesian word learning. In ACCSS, 2012.
- F. Xu and J.B. Tenenbaum. Word learning as Bayesian inference. Psychological Review, 114(2): 245-272, 2007.
- Y. Jia, J. Abbott, J. Austerweil, T. Griffiths, T. Darrell. Visually-Grounded Bayesian Word Learning. UC Berkeley EECS Tech Report 2012-202.