Learning Concept Taxonomies from Multi-modal Data

Hao Zhang

Zhiting Hu, Yuntian Deng, Mrinmaya Sachan, Zhicheng Yan and Eric P. Xing

Carnegie Mellon University

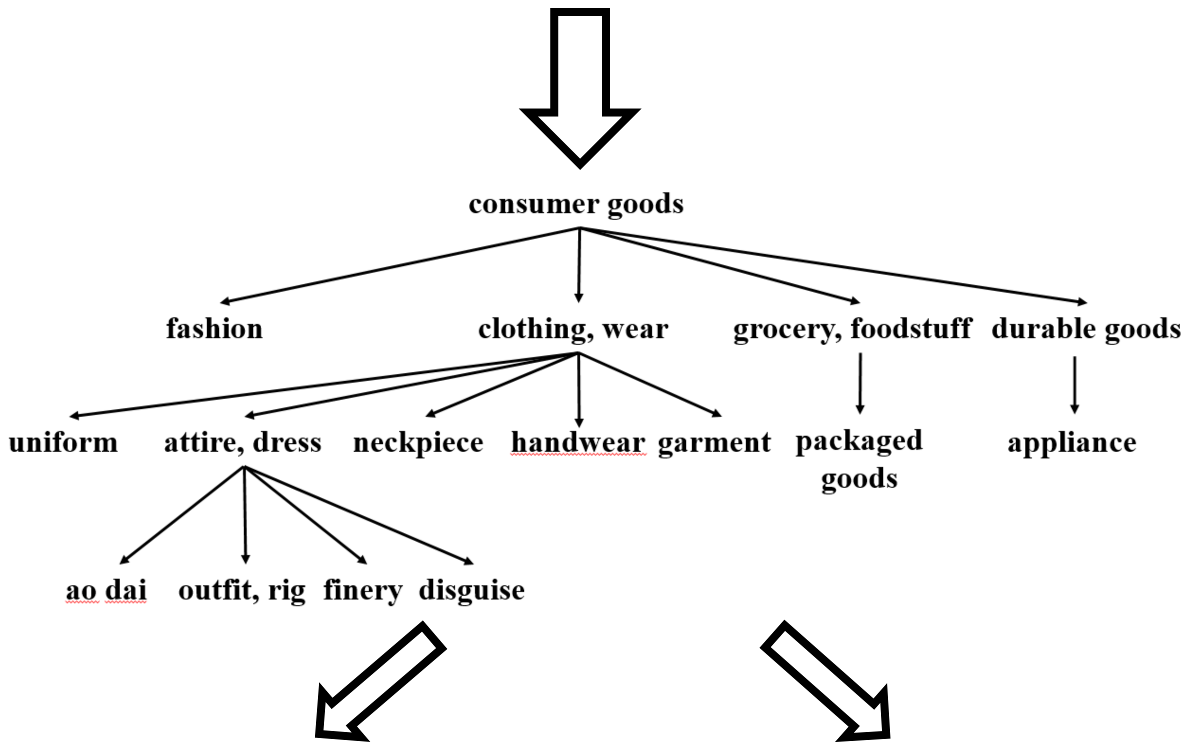
Outline

- Problem
- Taxonomy Induction Model
- Features
- Evaluation and Analysis

Problem

- Taxonomy induction

A set of lexical terms = {consumer goods, fashion, uniform, neckpiece, handwear, finery, disguise, ...}



- Human knowledge
- Interpretability

- Question answering
- Information extraction
- Computer vision

Problem

- Existing Taxonomies

WordNet

Thesaurus.com

IMAGENET

WIKIPEDIA
The Free Encyclopedia

  - Knowledge/time intensive to build
  - Limited coverage
  - Unavailable

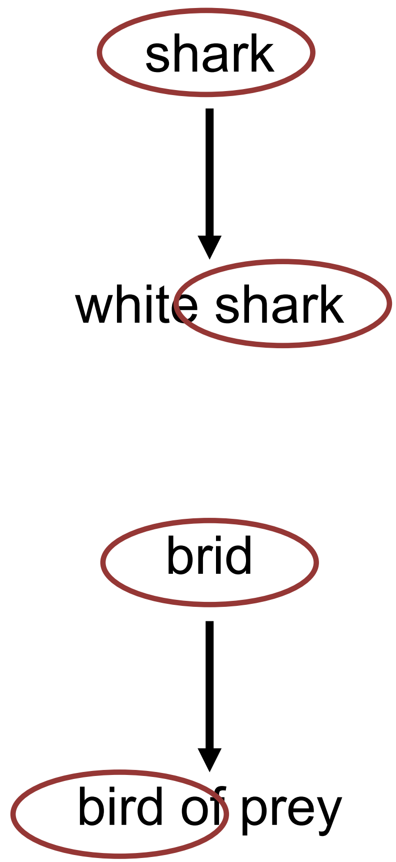
Related Works (NLP)

- Automatically induction of taxonomies

Widdows [2003]	**Snow et al [2006]**	**Poon and Domnigos [2010]**
Yang and Callan [2009]	**Kozareva and Hovy [2010]**	**Navigli et al [2011]**
Fu et al [2014]	**Bansal et al [2014]**	

Problem

- What evidence helps taxonomy induction?
  - Surface features
    - Ends with
    - Contains
    - Suffix match
    - …



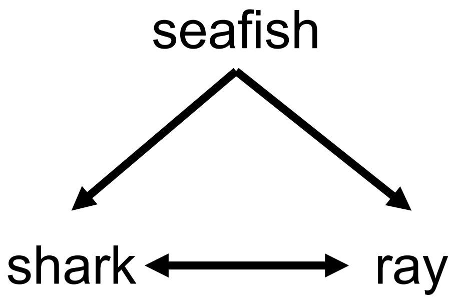
Problem

- What evidence helps taxonomy induction?
  - Semantics from text descriptions
    - Parent-child relation
    - Sibling relation [Bansal 2014]

“seafish, such as shark…”

“rays are a group
of seafishes…”

“Either shark or ray…”
“Both shark and ray…”



Problem

- What evidence helps taxonomy induction?
  - Semantics from text descriptions
    - Parent-child relation
    - Sibling relation [Bansal 2014]

"seafish, such as shark…"

"rays are a group
of seafishes…"

"Either shark or ray…"
"Both shark and ray…"

extracted as

- Wikipedia abstract
  - Presence and distance
  - Patterns
- Web-ngrams
- …

Problem

- What evidence helps taxonomy induction?
  - wordvec

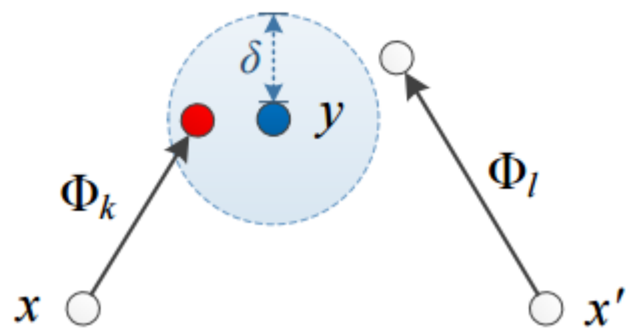
$$\mathrm{d}(v(king), v(queen)) \approx d(v(man), v(woman))$$

$$v(seafish) - v(shark) \overset{?}{\longleftrightarrow} v(human) - v(woman)$$

  - Projections between parent and child [Fu 2014]

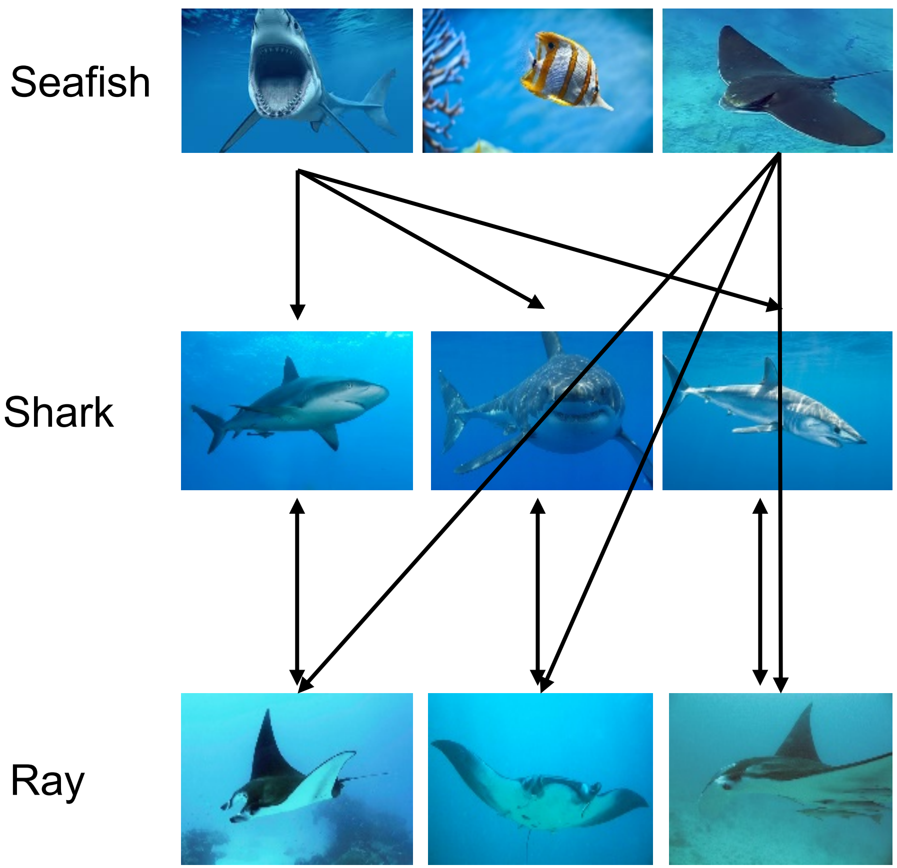
$$\Phi^* = \arg\min_{\Phi} \frac{1}{N} \sum_{(x,y)} \| \Phi x - y \|^2$$

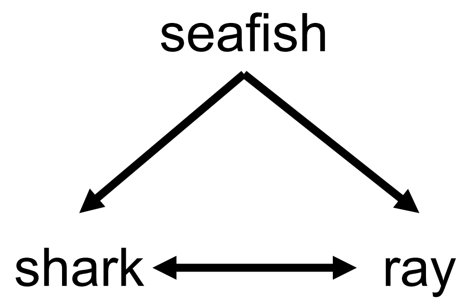
$$d(\Phi_k x, y) = \| \Phi_k x - y \|^2 < \delta$$



Motivation

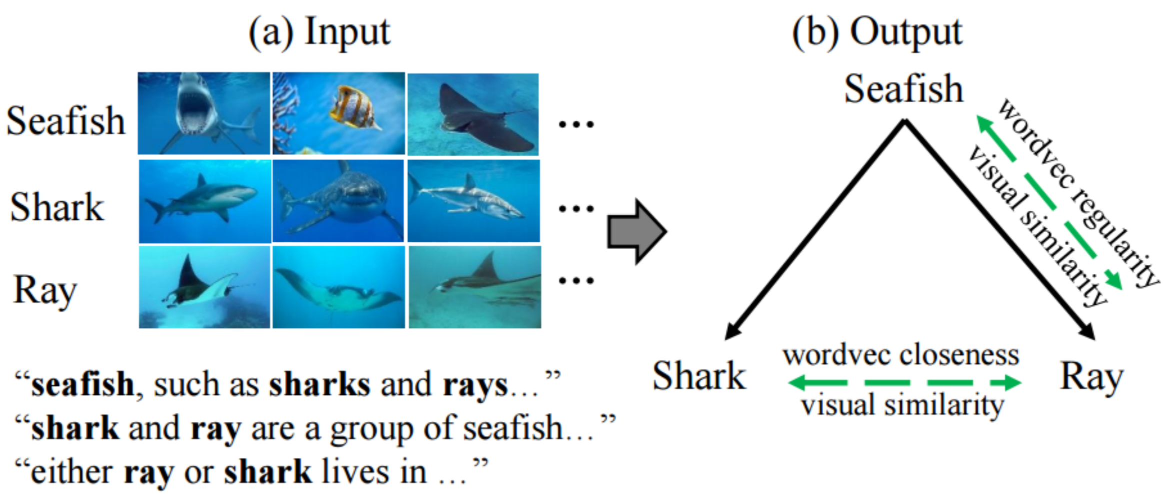
- How about images?





Motivation

- Our motivation
  - Images may include perceptual semantics
  - Jointly leverage text and visual information (from the web)



- Problems to be addressed:
  - How to design visual features to capture the perceptual semantics?
  - How to design models to integrate visual and text information?

Related Works (CV)

- Building visual hierarchies

Griffin and Perona [2008]

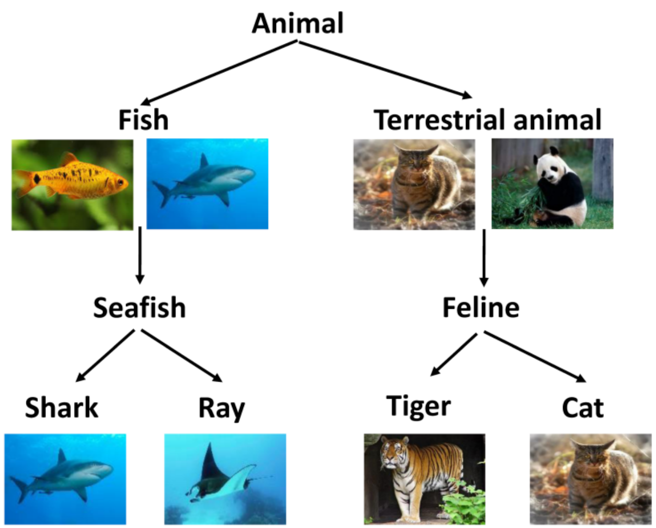
Sivic et al [2008]

Chen et al [2013]

Task Definition

- Assume a set of N cateogries $\boldsymbol{x} = \{x_1, x_2, \ldots, x_N\}$
  - Each category has a *name* and *a set of images*
- Goal: induce a taxonomy tree over $\boldsymbol{x}$
  - Using both text & visual features

$\boldsymbol{x}$ = {Animal, Fish, Shark, Cat, Tiger, Terrestrial animal, Seafish, Feline}



- Setting: Supervised learning of category hierarchies from data

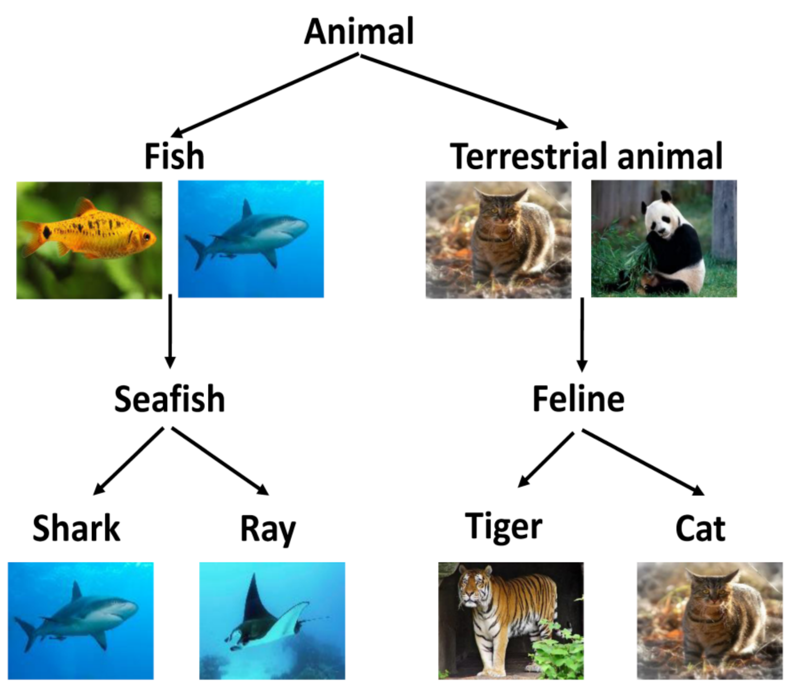
Model

Let $z_n (1 \leq z_n \leq N)$ be the index of the parent of category x_n

- The set $\mathbf{z} = \{z_1, z_2, \dots, z_n\}$ encodes the whole tree structure

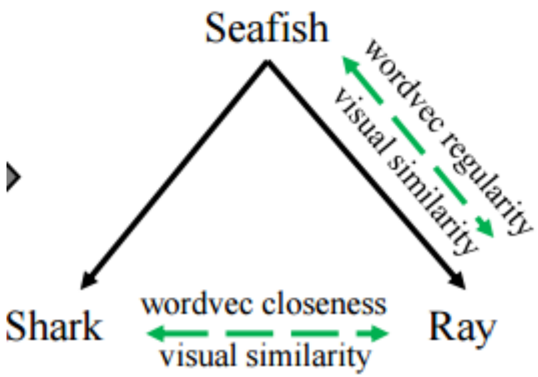
- Our goal → infer the conditional distribution $p(\boldsymbol{z}|\boldsymbol{x})$

x = {Animal, Fish, Shark, Cat, Tiger, Terrestrial animal, Seafish, Feline}



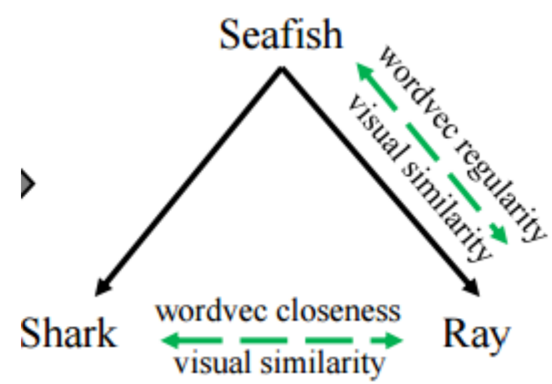
Model Overview

- Intuition: Categories tend to be closely related to **parents** and **siblings**
  - (text) hypernym-hyponym relation: *shark* -> *cat shark*
  - visual similarity: images of *shark* ⇔ images of *ray*
- Method: Induce features from **distributed representations** of images and text
  - image: deep convnet
  - text: word embedding



Taxonomy Induction Model

- Notations:
  - $\boldsymbol{c}_n$: child nodes of x_n
  - $x'_n \in \boldsymbol{c}_n$
  - g_w: consistency term depending on features
  - w: model weights to be learned



parent indexes of categories

popularity (#child) of categories

$$p_w(\boldsymbol{z}, \boldsymbol{\pi} | \boldsymbol{x}, \boldsymbol{\alpha}) \propto$$

$$p(\boldsymbol{\pi}|\alpha) \prod_{n=1}^{N} \prod_{x_{n'} \in \boldsymbol{c}_n} \pi_n g_w(x_n, x_{n'}, \boldsymbol{c}_n \backslash x_{n'}),$$

prior of popularity

consistency of <u>x'_n with parent x_n and siblings $\boldsymbol{c_n} \backslash \mathrm{x}_{n'}$</u>

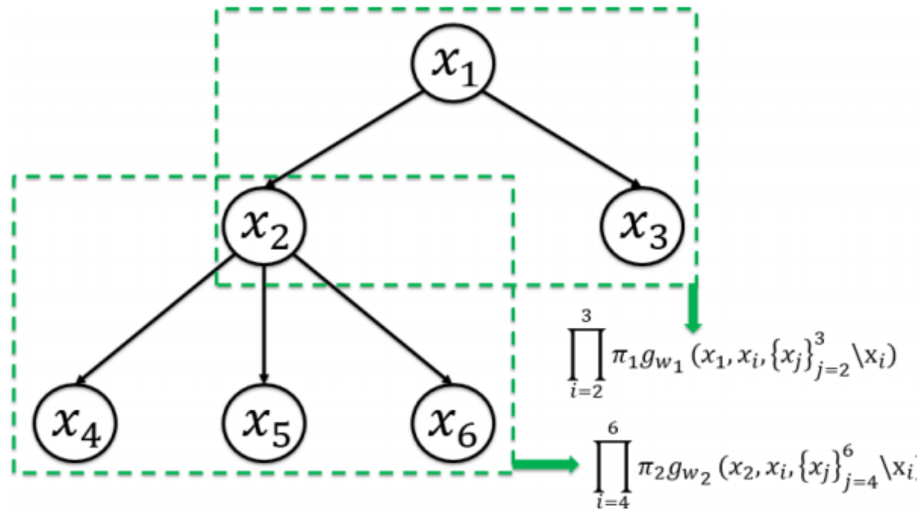
Taxonomy Induction Model

- Looking into g_w:
  - $g(x_n, x_n', \boldsymbol{c}_n \backslash x_n')$ evaluates how consistent a parent-child group is.
  - The whole model is a factorization of consistency terms of all local parent-child groups.

$$p_w(\boldsymbol{z}, \boldsymbol{\pi} | \boldsymbol{x}, \boldsymbol{\alpha}) \propto$$

$$p(\boldsymbol{\pi}|\alpha) \prod_{n=1}^{N} \prod_{x_{n'} \in \boldsymbol{c}_n} \pi_n g_w(x_n, x_{n'}, \boldsymbol{c}_n \backslash x_{n'}),$$

consistency of x_n' with parent x_n and siblings $\boldsymbol{c_n} \backslash x_{n'}$



Model: Develop g_w

- Notations:
  - $\boldsymbol{c}_n$: child nodes of x_n
  - $x_n' \in \boldsymbol{c}_n$
  - g_w: consistency term depending on features
  - w: model weights to be learned

weight vector (to be learned)

$$p_w(\boldsymbol{z}, \boldsymbol{\pi} | \boldsymbol{x}, \boldsymbol{\alpha}) \propto$$

$$p(\boldsymbol{\pi}|\alpha) \prod_{n=1}^{N} \prod_{x_{n'} \in \boldsymbol{c}_n} \pi_n g_w(x_n, x_{n'}, \boldsymbol{c}_n \backslash x_{n'}),$$

$$\exp\left\{\boldsymbol{w}^\top \boldsymbol{f}_{n,n',\boldsymbol{c}_n \backslash x_{n'}}\right\}$$

consistency of x_n' with parent x_n and siblings $\boldsymbol{c}_n \backslash x_{n'}$

feature vector: feature vector of x_n' with parent x_n and siblings $\boldsymbol{c}_n \backslash x_n'$

Feature: Develop f

- Visual features:
  - Sibling similarity
  - Parent-child similarity
  - Parent prediction
- Text features
  - Parent prediction [Fu et al.]
  - Sibling Similarity
  - Surface features [Bansal et al.]

Feature: Develop f

- Visual features: Sibling similarity (S-V1*)
  - Step 1 : fit a Gaussian to the images of each category
  - Step 2: Derive the **pairwise** similarity $vissim(x_n, x_m)$

$$vissim(x_n, x_m) = [\mathcal{N}(\overline{\boldsymbol{v}}_{\boldsymbol{i}_m}; \overline{\boldsymbol{v}}_{\boldsymbol{i}_n}, \Sigma_n) + \mathcal{N}(\overline{\boldsymbol{v}}_{\boldsymbol{i}_n}; \overline{\boldsymbol{v}}_{\boldsymbol{i}_m}, \Sigma_m)]/2$$

  - Step 3: Derive the **groupwise** similarity by averaging

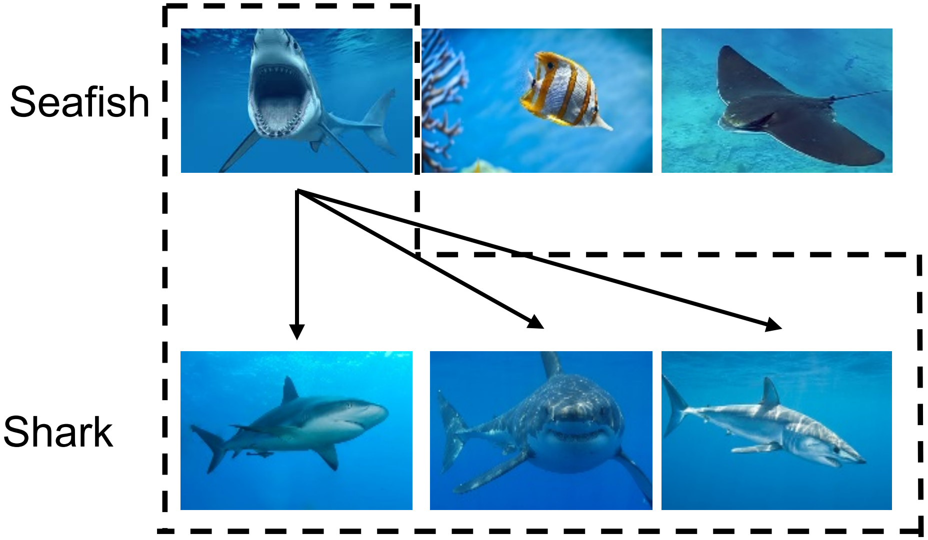
$$vissim(x_{n'}, \boldsymbol{c}_n \backslash x_{n'}) = \frac{\sum_{x_m \in \boldsymbol{c}_n \backslash x_{n'}} vissim(x_{n'}, x_m)}{|\boldsymbol{c}_n| - 1}.$$

S-V1 evaluates the visual similarity between siblings

* S: Siblings, V: Visual

Feature: Develop f

- Visual features: Parent-child Similarity (PC-V1*)
  - Step 1 : Fit a Gaussian for child categories
  - Step 2: Fit a Gaussian for **only the top-K images of parent categories**
  - Step 3 – 4: same with S-V1



* PC: Parent-child, V: Visual

Feature: Develop f

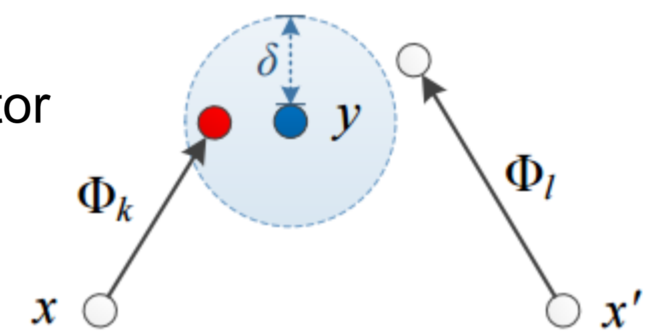
- Visual features: Parent Prediction (PC-V2*)
  - Step 1 : Learn a projection matrix to map the mean image of child category to the word embedding of its parent category

$$\boldsymbol{\Phi}^* = \underset{\boldsymbol{\Phi}}{\operatorname{argmin}} \frac{1}{N} \sum_n \|\boldsymbol{\Phi}\overline{\boldsymbol{v}}_{\boldsymbol{i}_{n'}} - \boldsymbol{v}_{t_n}\|_2^2 + \lambda \|\boldsymbol{\Phi}\|_1$$

  - Step 2: Calculate the distance

$$\|\boldsymbol{\Phi}\overline{\boldsymbol{v}}_{\boldsymbol{i}_{n'}} - \boldsymbol{v}_{t_n}\|$$

  - Step 3: bin the distance as a feature vector



* PC: Parent-child, V: Visual

Feature: Develop f

- Text features
  - Parent prediction [Fu et al.]
    - Parent prediction: projection from child to parent
  - Sibling Similarity
    - Distance between word vectors
  - Surface features [Bansal et al.]
    - Ends with (e.g. catshark is a sub-category of shark), LCS, Capitalization, etc.

Parameter Estimation

- Inference
  - Gibbs sampling

$$p(z_n = m | \boldsymbol{z} \backslash z_n, \cdot)$$
$$\propto \left(q_m^{-n} + \alpha_m\right) \frac{\prod_{x_{n'} \in \boldsymbol{c}_m \cup \{x_n\}} g_w\left(x_m, x_{n'}, \boldsymbol{c}_m \cup \{x_n\}\right)}{\prod_{x_{n'} \in \boldsymbol{c}_m \backslash x_n} g_w\left(x_m, x_{n'}, \boldsymbol{c}_m \backslash x_n\right)}$$

- Learning
  - Supervised learning from gold taxonomies of training data
  - Gradient descent-based maximum likelihood estimation
- Output taxonomies
  - Chao-Liu-Edmonds algorithm

Experiment Setup

- Implementation
  - Wordvec: Google word2vec
  - Convnet: VGG-16
- Evaluation metric: $\text{Ancestor-F1} = \frac{2PR}{P+R}$

$$P = \frac{|\text{is-a}_{\text{predicted}}| \cap |\text{is-a}_{\text{gold}}|}{|\text{is-a}_{\text{predicted}}|}, R = \frac{|\text{is-a}_{\text{predicted}}| \cap |\text{is-a}_{\text{gold}}|}{|\text{is-a}_{\text{gold}}|}$$

- Data
  - Training set: ImageNet taxonomies

Trees	Tree A	Tree B	Tree C
Synset ID	12638	19919	23733
Name	consumer goods	animal	food, nutrient
$h = 4$	187	207	572
$h = 5$	362	415	890
$h = 6$	493	800	1166
$h = 7$	524	1386	1326

Evaluation

Results: Comparison to baseline methods

- Embedding-based feature (LV) is comparable to state-of-the-art
- Full feature set (LVB) achieve the best

Method	$h = 4$	$h = 5$	$h = 6$	$h = 7$
Hierarchy Completion				
Fu2014	0.66	0.42	0.26	0.21
Ours (L)	0.70	0.49	0.45	0.37
Ours (LV)	**0.73**	**0.51**	**0.50**	**0.42**
Hierarchy Construction				
Fu2014	0.53	0.33	0.28	0.18
Bansal2014	0.67	0.53	0.43	0.37
Ours (L)	0.58	0.41	0.36	0.30
Ours (LB)	0.68	0.55	0.45	0.40
Ours (LV)	0.66	0.52	0.42	0.34
Ours (LVB - E)	0.68	0.55	0.44	0.39
Ours (LVB)	**0.70**	**0.57**	**0.49**	**0.43**

- L: **L**anguage features
  - surface features
  - embedding features
- V: **V**isual features
- B: **B**ansal2014 features
  - web ngrams etc.
- E: **E**mbedding features

Evaluation

Results: How much visual features help?

S-V1	PC-V1	PC-V2	h = 4	h = 5	h = 6	h = 7
			0.58	0.41	0.36	0.30
✓			0.63	0.48	0.40	0.32
	✓		0.61	0.44	0.38	0.31
		✓	0.60	0.42	0.37	0.31
✓	✓		0.65	**0.52**	0.41	0.33
✓	✓	✓	**0.66**	**0.52**	**0.42**	**0.34**

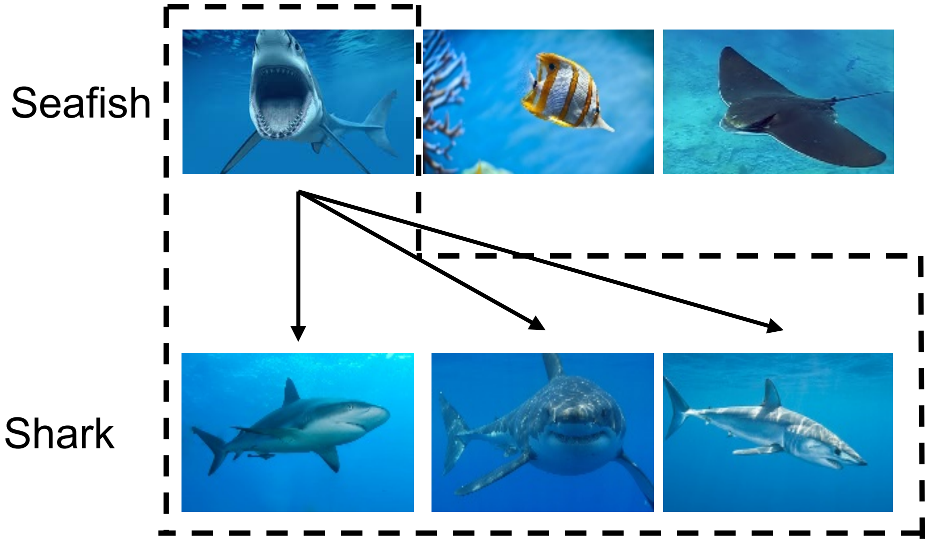
Messages:

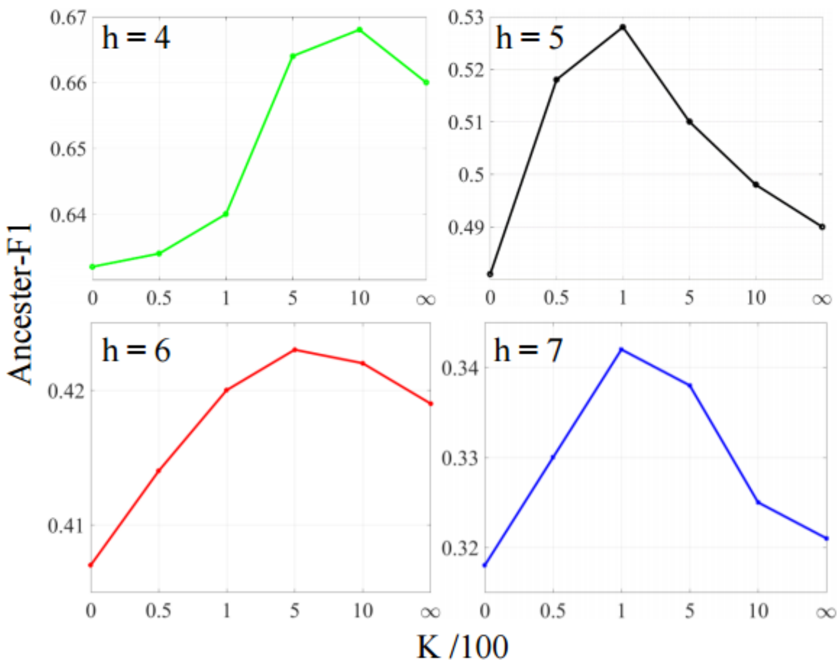
- Visual similarity (S-V1, PC-V1) help a lot
- The complexity of visual representations does not affect much

Evaluation

Results: Investigating PC-V1

- Images of parent category are not all necessarily visually similar to images of child category



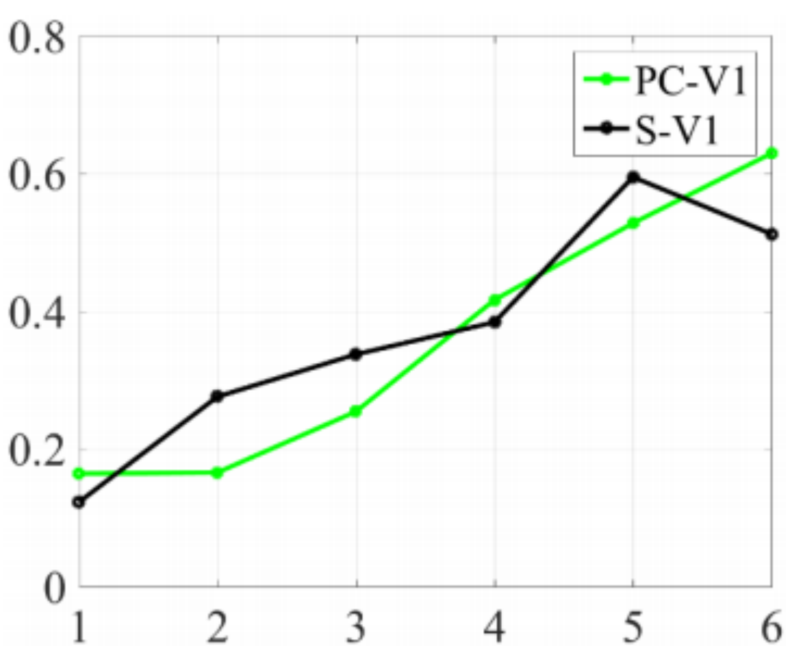


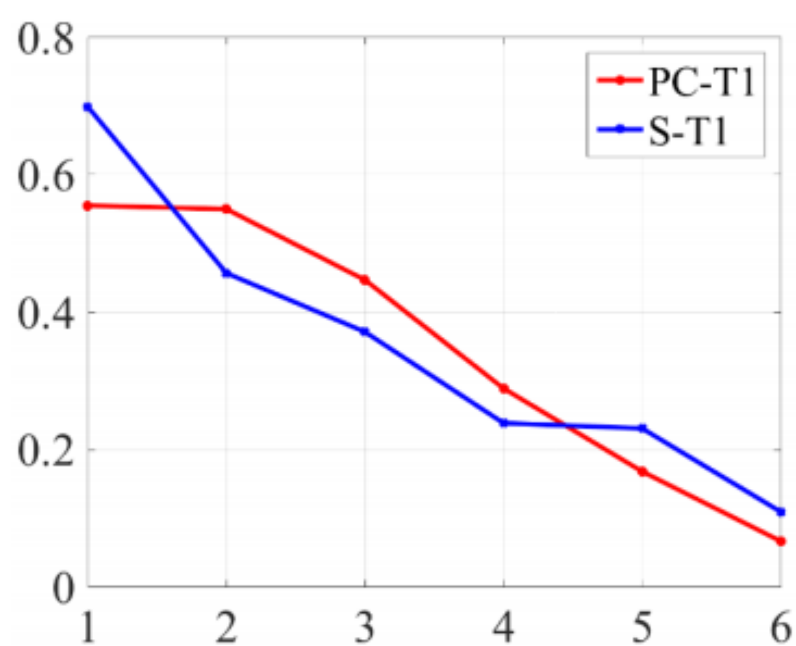
Evaluation

Results: When/Where visual features help?

- Messages:
  - Shallow layers ↔abstract categories ↔ text features more effective
  - Deep layers ↔ specific categories ↔ visual features more effective

Weights v.s. depth





Take-home Message

- Visual similarity helps taxonomy induction a lot
  - Sibling similarity
  - Parent-child similarity
- Which features are more important?
  - Visual features are more indicative in near-leaf layers
  - Text features more evident in near-root layers
- Embedding features augments word count features

Thank You!

Q & A

Evaluation

Results: Visualization