

CHOOSING BASIC-LEVEL CONCEPT NAMES USING VISUAL AND LANGUAGE CONTEXT

Alexander Mathews, Lexing Xie and Xuming He Australian National University, NICTA

GOAL AND CONTRIBUTIONS

We predict which words people will associate with an image, using three main ideas:

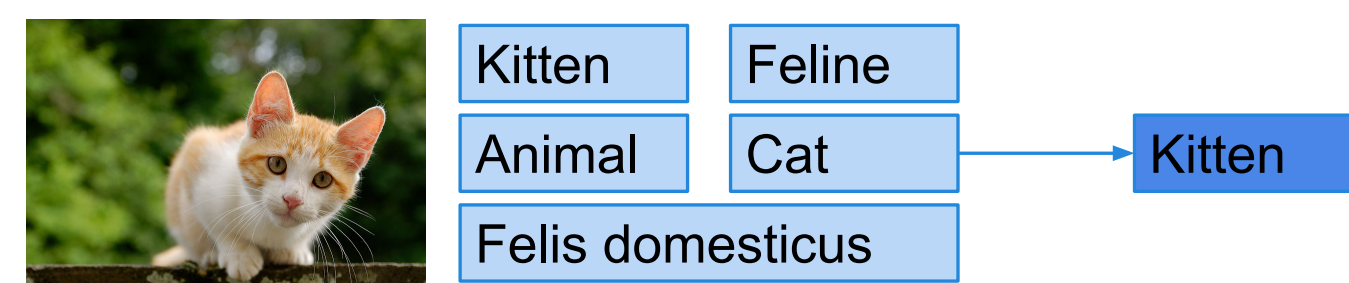
- Implement *basic-level categories* from cognitive psychology;
- Use visual context such as object and scene properties;
- Model language context such as statistical co-occurrence of words.

There are three key contributions:

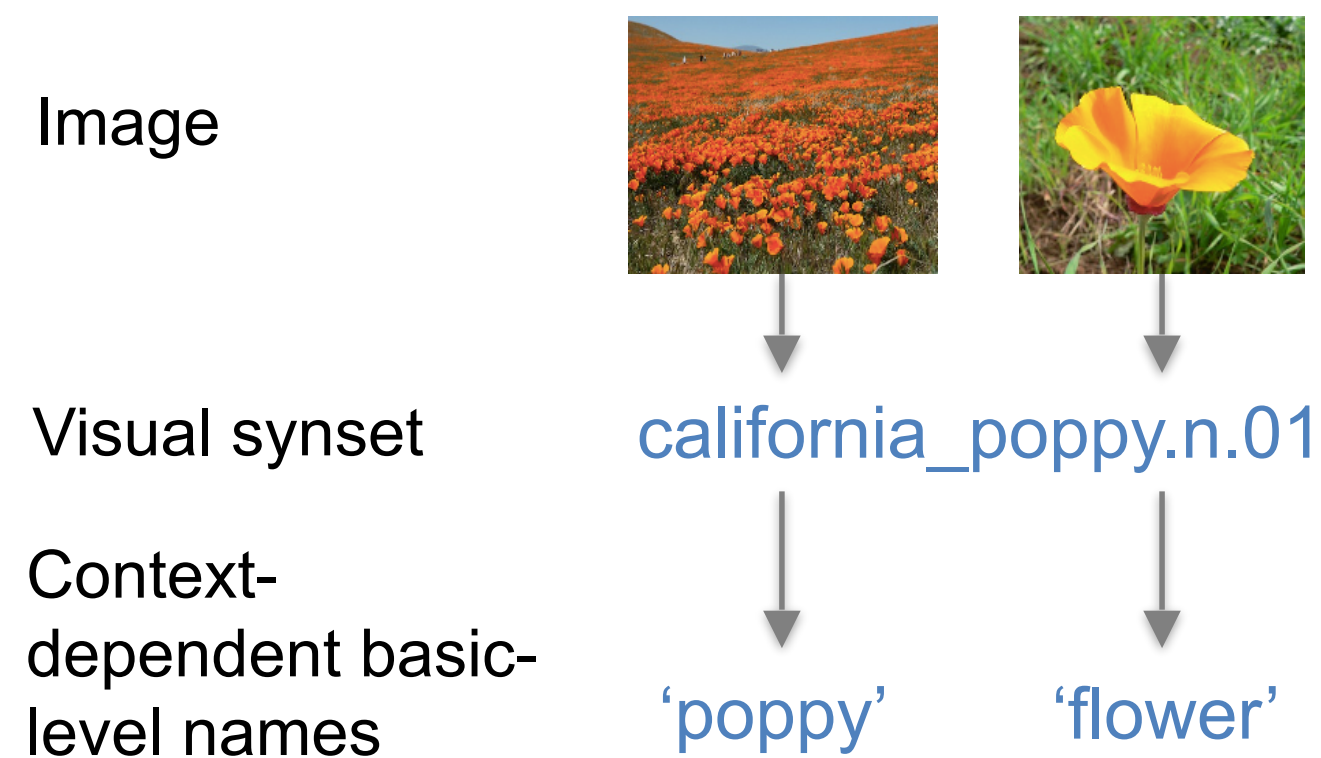
- A new method to predict context-dependent basic-level categories.
- The first large-scale catalogue of context-dependent basic-level categories, of thousands of visual concepts and hundreds of thousands of images.
- A word ranking benchmark on a dataset two orders of magnitude larger than in previous work [1], with consistent improvements.

BASIC-LEVEL CATEGORIES

The *basic* level of categorization is “the most inclusive (abstract) level at which the categories can mirror the structure of attributes perceived in the world” [2].



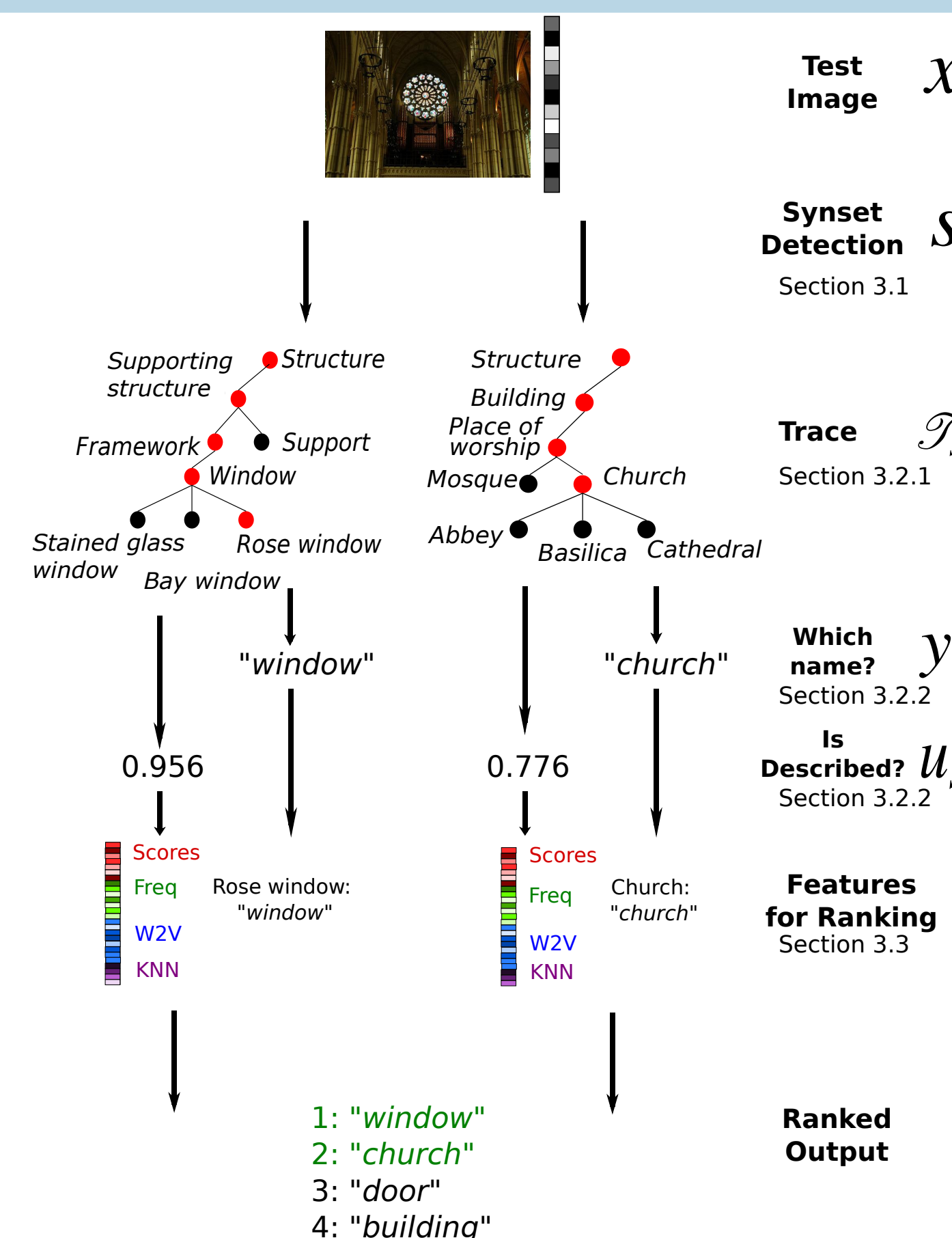
Basic-level categories is influenced by context – people are known to choose different names for visual concepts depending on (a) visual attributes of the object, (b) contextual priming, and (c) the rest of the visual scene [2].



METHOD OVERVIEW

Given image \mathbf{x}_i , our system predicts the most likely words in three steps:

- (1) Detect synset s .
- (2) Identify the most likely basic-level names for synset s , by computing the probability of each possible *name* y^s from a candidate set \mathcal{T}_s , and whether or not synset s is described (u_s).
- (3) Rank all names y_k for all concepts s_m in image i , by computing a rank score $r_{i,m,k}$.



METHOD

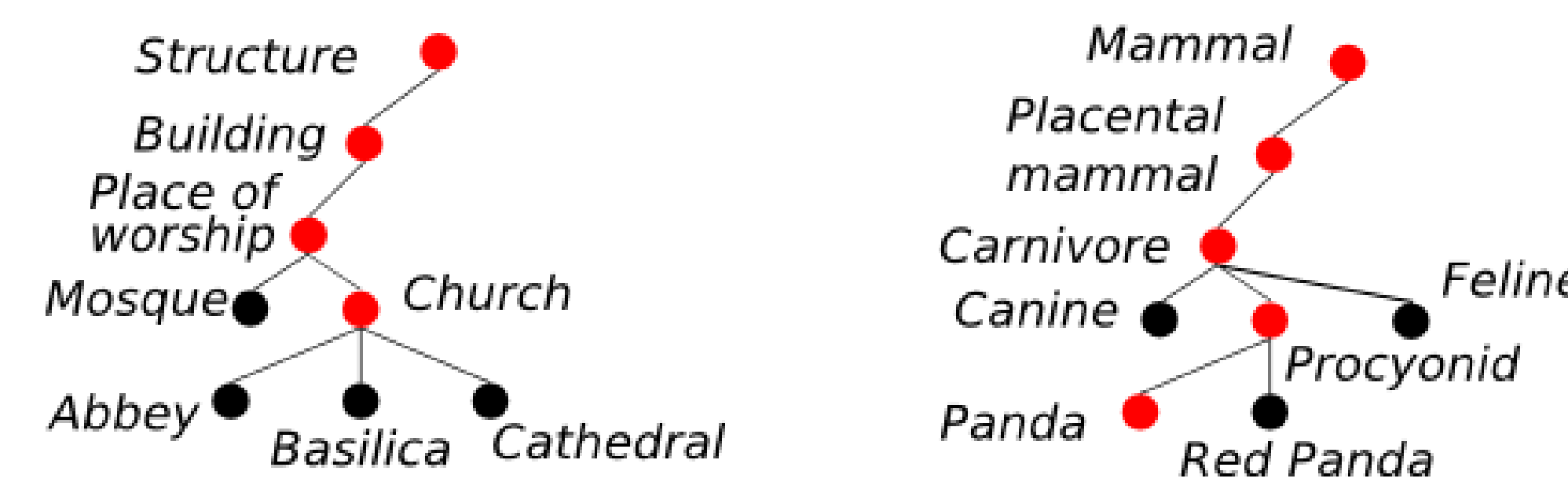
Detecting visual concepts

$$p(s = 1|\mathbf{x}) = \sigma(\mathbf{w}_s^T \mathbf{x}) \quad (1)$$

2633 concept detectors are trained with ImageNet dataset by adapting the last supervised layer of a Convolutional Neural Network.

Generating basic-level name candidates

- Tracing the WordNet hierarchy up 5 levels;
- extracting the lemmas of each ancestor synset.



Choosing basic-level names

$$p(y_i^s = 1|\mathbf{x}, s) = \sum_{u_s \in \{0,1\}} p(y_i^s = 1|\mathbf{x}, s, u_s) p(u_s|\mathbf{x}, s) \\ = p(y_i^s = 1|\mathbf{x}, s, u_s = 1) p(u_s = 1|\mathbf{x}, s) \quad (2)$$

For each visual concept s ,

- Learn $p(u_s = 1|\mathbf{x}, s)$, the probability that synset s is described.
- Learn $p(y_i^s = 1|\mathbf{x}, s, u_s = 1)$, to choose among the possible names.

Synset	mTurk	Ngram	Description Classifier	Synset	Description Classifier
boatbill .n.01	bird	bird	bird	cathedral .n.02	building
			heron		cathedral
white_ash .n.01	leaf	ash	plant	church	church
			tree		art
minivan .n.01	van	van	car	sculpture	sculpture
			van		carving

Ranking basic-level names across synsets

$$r_{i,m,k} = \mathbf{w}_r^T h_{i,m,k}; \quad r_{i,m,k} > r_{i,q,l} \quad (3)$$

We learn a linear ranking function using a ranking objective that prefers synset m name k that appeared with image i over synset q name l that did not appear with the same image. The ranking features include:

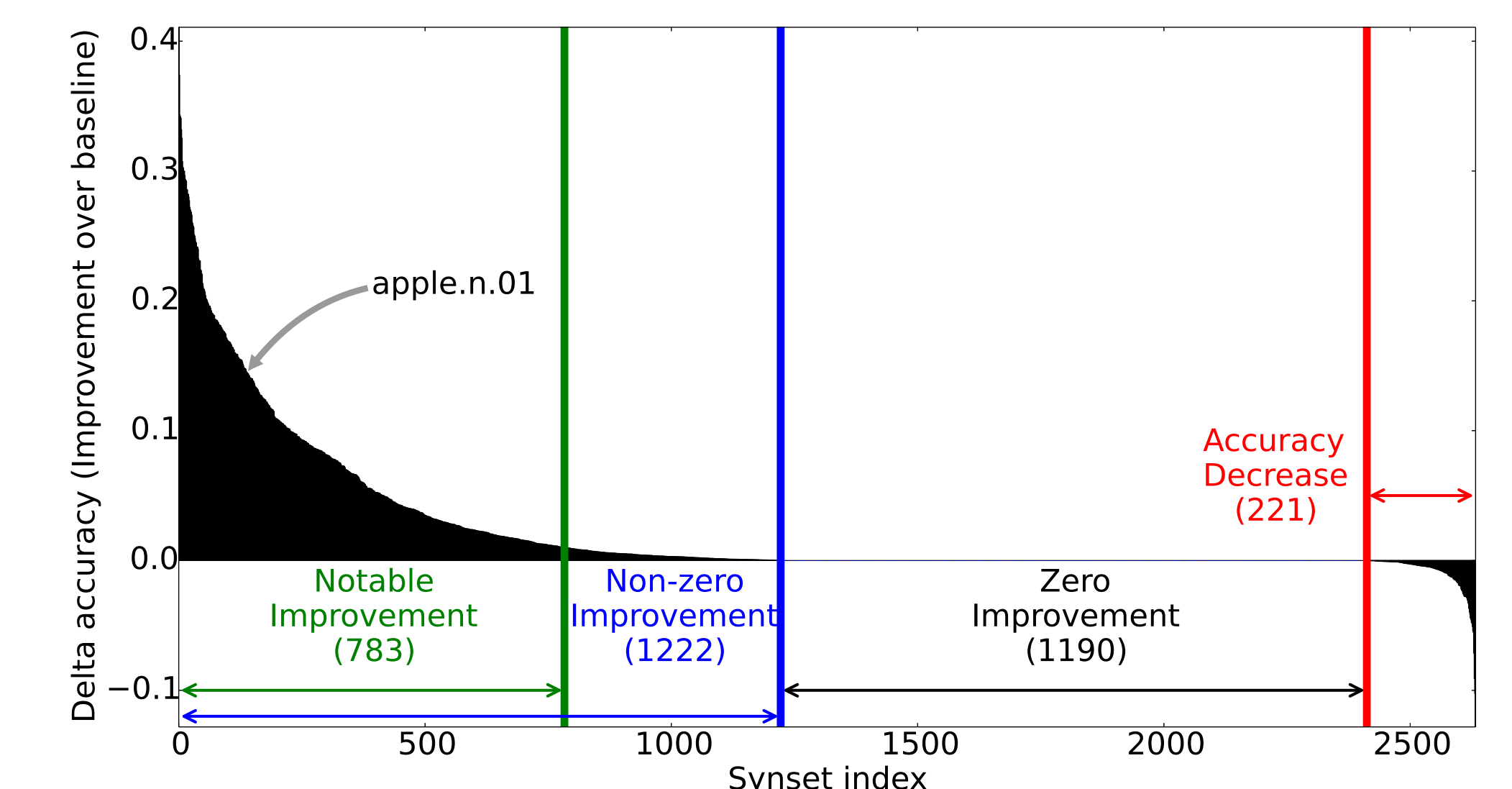
- SCORES from classifiers at different stages;
- AUX-liary information about classifiers and classification targets;
- KNN – nearest images with TF-IDF over their captions;
- WORD2VEC features consisting of the vector-space similarity and probability of the target word given other context words.

EXPERIMENTAL RESULTS

Datasets

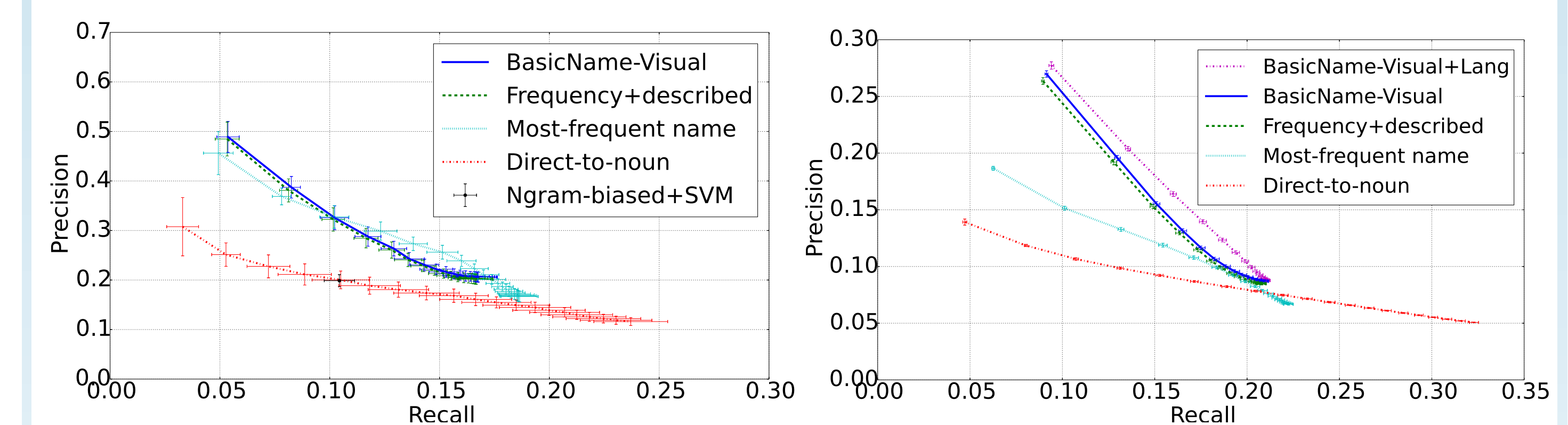
- IMAGENET-FLICKR: Training synset classifiers.
- 80% of SBU-1M images: Training basic-level name classifiers.
- SBU-1K A and B: Evaluation with words generated by MTurk.
- SBU-148K: Evaluation with words from Flickr captions.

Accuracy improvements of basic-level name classification over the *Frequency+described* baseline for 2,633 synsets. The percentage of improved synsets is on par with the percentage of synsets with ambiguous basic-level names – two or more names used with similar frequency.



Precision-recall curves on SBU-1KA (left) and SBU-148K (right).

- Our methods: *BasicName-Visual* and *BasicName-Visual+Lang*
- Four baselines: varying amounts of naming information.



Word ranking result on example images.

Images	Labels	Ngram-biased-SVM	Direct-to-noun	Frequency+described	BasicName-Visual+Lang
	altar, alter, roof, art, building, flower, church, door, lamp, light, wall, podium, window, stained glass, vase, stain glass window,	street building door tower room	tree window building tower monument column	window building tower monument column	window building tower church monument
	close-up, flower, petal, sky, tree, stamen, sunflower	bird pink tree plant white	house girl sign dog mountain	flower plant yellow	sunflower flower yellow plant
	bubble, float, lake, man, pants, plastic, pond, shirt, shrub, water	shift dog zoo grass ball	water girl river beach house	ball fish building sail bird	ball fish building way sail

References

- [1] V. Ordonez, J. Deng, Y. Choi, A. Berg and T. Berg From large scale image categorization to entry-level categories ICCV, 2013.
- [2] E. Rosch, Principles of categorization. Concepts: core readings, 1999.