**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**

**Detecting visual concepts**

\[ p(s = 1|x) = \sigma(w^T \cdot x) \]  \hspace{1cm} (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

\[ p(y^*_i = 1|x, s) = \sum_{u_i \in \{0, 1\}} p(y^*_i = 1|x, s, u_i)p(u_i|x, s) \]  \hspace{1cm} (2)

For each visual concept \( s \),
- Learn \( p(u_i = 1|x, s) \), the probability that synset \( s \) is-described.
- Learn \( p(y^*_i = 1|x, s, u_i = 1) \), to choose among the possible names.

**Ranking basic-level names across synsets**

\[ r_{i, m, k} = w^T \cdot h_{i, m, k} \]  \hspace{1cm} (3)

We learn a linear ranking function using a ranking objective that prefers synset \( n \) 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-iliary 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-1KA and SBU-1KB: 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**.

**References**