Sense Discovery via Co-Clustering on Images and Text

Xinlei Chen[†] Alan Ritter[‡] Abhinav Gupta[†] Tom Mitchell[†]
[†]Carnegie Mellon University [‡]Ohio State University

How many different concepts can the Noun Phrase (NP) *Columbia* refer to? In this paper, we present an unsupervised approach for discovering multiple senses of a given NP (Figure 1). Our algorithm not only discovers multiple semantic senses but also multiple visual senses of a given NP. For example, the NP *Columbia* can refer to the *university*, the *sportswear company*, the *river* and the *Studio*. Similarly, in the visual world, *Columbia* can refer to images of the university building, professor, shoes, jacket and even the scenic columbia river.

Sense discovery is an important problem originating from natural language and knowledge representation. Recently, with a growing interest in building large-scale visual knowledge bases automatically [1, 3, 4], this problem is attracting increasing attention since a good knowledge base must first disambiguate different meanings before learning the relationships between different concepts. One obvious way to handle this problem is to fall back on human-developed knowledge bases such as Wordnet [3]. While the resulting senses are clean, this approach lacks scalability and coverage. In this paper, instead of relying on manually-compiled resources, we focus on a data-driven approach to discover multiple senses of a given NP in an unsupervised manner.

The most common approach for sense discovery is to represent each instance of a NP in terms of text and/or image features and then cluster these instances to obtain multiple semantic and visual senses of the NP respectively. Most joint clustering approaches make the simplifying assumption that there exists a one-to-one mapping between semantic and visual senses of a word. However, this assumption rarely holds in practice. For example, while there are two predominant semantic senses of the NP *Apple*, there exist multiple visual senses due to appearance variation (green *vs*. red apples), viewpoint changes, *etc*.

Based on this key observation, we propose a generalized co-clustering algorithm for this cross-domain sense disambiguation problem. Unlike traditional clustering approaches which assume a one-to-one mapping between the clusters in the text-based feature space and the visual space, we adopt a one-to-many mapping between the two spaces. We use this one-to-many mapping to provide constraints on the co-clustering algorithm. This not only leads to high purity clusters, but more importantly, this joint learning process allows us to infer an alignment between the semantic and visual senses of the NP. Specifically, given image-text pairs for a given NP (obtained using search engines), we first use an exemplar based approached proposed in [2] to discover multiple visual senses in the image domain, while keeping a single semantic cluster on the text side. This is followed by a structure EM-like approach that iteratively: (a) labels instances and clusters them separately in the two feature space; (b) builds image and text classifiers for each cluster to clean up the data; (c) learns a mapping between the clusters.

Extensive experiments are done with data readily available from the web. First, we introduce a new challenging dataset for this task (CMU Polysemy-30). This dataset consists of 30 NPs and ∼750 image-text pairs from Google Image Search. We manually labeled ∼5600 instances with one of the listed semantic senses. To overcome the human labeling bottleneck, we also perform another experiment which creates pseudo-words to evaluate sense disambiguation. Next, we perform a retrieval experiment on the MIT ISD dataset [5] which has five polysemous NPs. Finally, we perform a large-scale experiment where we ran our algorithm on ∼2000 NPs. The list of these concepts is obtained using the NEIL knowledge base [1]. It is shown that our algorithm is effective in not only discovering multiple senses and clustering the data but it also generates the right mapping between semantic and visual senses.

[1] Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. Neil: Extracting visual knowledge from web data. In *ICCV*, 2013.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.
http://www.neil-kb.com/poly.html

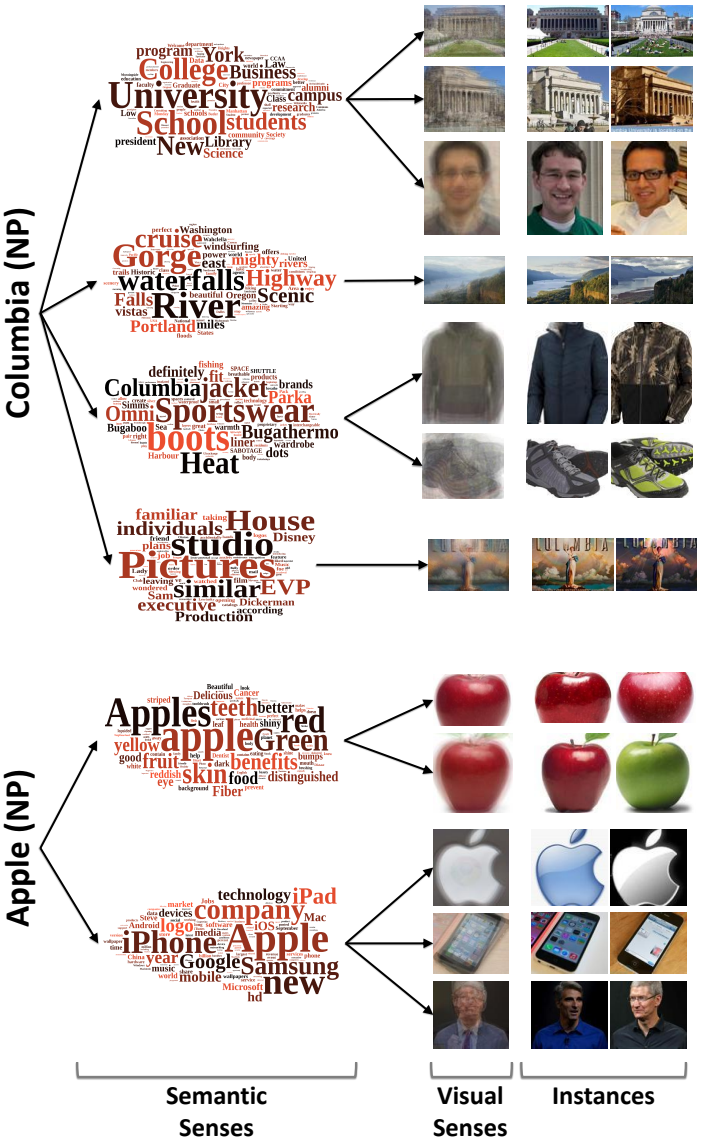

Figure 1: We present a co-clustering algorithm that discovers multiple semantic and visual senses of a given NP. In the figure above, we show the multiple senses discovered for *Columbia* and *Apple*. In the case of *Columbia*, our approach automatically discovers four semantic senses: *university*, *river*, *sportswear*, *studio*. In case of *Apple*, it discovers two semantic senses: *fruit*, *company*. Our approach also discovers multiple visual senses. For example, the *sportswear* sense of *Columbia* corresponds to two visual senses: *jacket* and *shoes*. Semantic senses are shown as word clouds with size of each word being proportional to its importance. Visual senses are shown as average images of members belonging to the cluster.

[2] Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. Enriching visual knowledge bases via object discovery and segmentation. In *CVPR*, 2014.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

[4] Santosh K Divvala, Ali Farhadi, and Carlos Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *CVPR*, 2014.

[5] Kate Saenko and Trevor Darrell. Unsupervised learning of visual sense models for polysemous words. In *NIPS*, 2008.