Stating the Obvious:
Extracting Visual Common Sense Knowledge

Mark Yatskar\textsuperscript{1}, Vicente Ordonez\textsuperscript{2}, Ali Farhadi\textsuperscript{1,2}
\textsuperscript{1}Computer Science & Engineering, University of Washington, Seattle, WA
\textsuperscript{2}Allen Institute for Artificial Intelligence (AI2), Seattle, WA
[my89, ali]@cs.washington.edu, vicenteor@allenai.org

Abstract

Obtaining common sense knowledge using current information extraction techniques is extremely challenging. In this work, we instead propose to derive simple common sense statements from fully annotated object detection corpora such as the Microsoft Common Objects in Context dataset. We show that many thousands of common sense facts can be extracted from such corpora at high quality. Furthermore, using WordNet and a novel submodular $k$-coverage formulation, we are able to generalize our initial set of common sense assertions to unseen objects and uncover over 400k potentially useful facts.

1 Introduction

How can we discover that bowls can hold broccoli, that if a knife touches a cake then a person is probably cutting cake, or that cutlery can be on dining tables? We propose to leverage the effort of computer vision researchers in creating large scale datasets for object detection and use these resources instead to extract symbolic representations of visual common sense. The knowledge we compile is physical, not commonly covered in text and more exhaustive than what people can usually produce.

Our focus is particularly on visual common sense, defined as the information about spatial and functional properties of entities in the world. We propose to extract three types of knowledge from the Microsoft Common Objects in Context dataset (Lin et al., 2014) (MS-COCO), consisting of 300,000 images, covering 80 objects, with object segments and natural language captions. First, we find spatial relations, e.g. $\text{hold}(\text{bed}, \text{dog})$, from outlines of co-occurring objects. Next, we construct entailment rules like $\text{hold}(\text{bed}, \text{dog}) \Rightarrow \text{laying-on}(\text{dog, bed})$ by associating spatial relations with text in captions. Finally, we uncover general facts such as $\text{hold}(\text{furniture, domestic animal})$, applicable to object types not present in MS-COCO by using WordNet (Miller, 1995) and a novel submodular $k$-coverage formulation.

Evaluations using crowdsourcing show our methods can discover many thousands of high quality explicit statements of visual common sense. While some of this knowledge can be potentially extracted from text (Vanderwende, 2005), we found that from our top 100 extracted spatial relations, e.g. $\text{hold}(\text{bed, dog})$, only 4 are present in some form in the AtLocation relations in the popular ConceptNet (Speer and Havasi, 2013) knowledge base. This shows that the knowledge we derive provides complimentary information for other more general knowledge bases. Such common sense facts have proved useful for query expansion (Kotov and Zhai, 2012; Bouchoucha et al., 2013) and could benefit entailment (Dagan et al., 2010), grounded entailment (Bowman et al., 2015), or visual recognition tasks (Zhu et al., 2014).

2 Related Work

Common sense knowledge has been predominately created directly from human input or extracted from text (Lenat et al., 1990; Liu and Singh, 2004; Carlson et al., 2010). In contrast, our work is focused on visual common sense extracted from images anno-
Our objective in this section is to score and rank a set of relations $S_1 = \{r(o_1, o_2)\}$, where $r$ is an object-object relation and $o_1, o_2 \in V$, using a function $\gamma_1 : S_1 \rightarrow \mathbb{R}$. First, we define a vocabulary $R$ of object-object relations between pairs of annotated objects. Our relations are inspired by Region Connection Calculus (Randell et al., 1992), and the Visual Dependency Grammar of (Elliott et al., 2014; Elliott and de Vries, 2015), details in Figure 1.

For every image, we record the instances of each of these object-object relations $r(o_1, o_2)$ between all co-occurring objects in $D^1$. We use Point-wise Mutual Information (PMI) to estimate the evidence for

---

1For symmetric relations like $above(o_1, o_2)$, and $under(o_1, o_2)$ we only record one of the relations.

---

3 Methods

We assume the availability of an object-level annotated image dataset $D$ containing a set of images with textual descriptions. Each object in an image must be annotated with: (1) a mask or polygon outlining the extents of the object, and (2) the category of the object from a set of categories $V$ and (3) an overall description of the image.

We produce three types of common sense facts, each with an associated scoring function: (1) Object-object relationships implicitly encoded in the relative configurations between objects in the annotated image data, i.e. $on(\text{bed}, \text{dog})$ (sec 3.1), (2) Entailment relations encoded in the relationships between object-object configurations and textual descriptions i.e. $on(\text{bed}, \text{dog}) \Rightarrow laying-on(\text{bed}, \text{dog})$ (sec 3.2), and (3) Generalized relations induced by using the semantic hierarchy of concepts in WordNet, i.e. $on(\text{furniture}, \text{domestic-animal})$ (sec 3.3).

3.1 Mining Object-Object Relations

Our objective in this section is to score and rank a set of relations $S_1 = \{r(o_1, o_2)\}$, where $r$ is a object-object relation and $o_1, o_2 \in V$, using a function $\gamma_1 : S_1 \rightarrow \mathbb{R}$. First, we define a vocabulary $R$ of object-object relations between pairs of annotated objects. Our relations are inspired by Region Connection Calculus (Randell et al., 1992), and the Visual Dependency Grammar of (Elliott et al., 2014; Elliott and de Vries, 2015), details in Figure 1.

For every image, we record the instances of each of these object-object relations $r(o_1, o_2)$ between all co-occurring objects in $D^1$. We use Point-wise Mutual Information (PMI) to estimate the evidence for

---

1For symmetric relations like $above(o_1, o_2)$, and $under(o_1, o_2)$ we only record one of the relations.
each relationship triplet:
\[
\gamma_1(r(o_1, o_2)) = \log \frac{p[r(o_1, o_2)]}{p[r]p(o_1, o_2)}, \tag{1}
\]

We estimate these probabilities by counting object-object-relation co-occurrences using existential quantifiers for every image. This means every image can at most contribute one to the count of \(r(o_1, o_2)\) so that we do not exacerbate the results by images with many identical object types taken from unusual viewpoints. In Figure 2, we provide examples of our extracted object-object relations.

### 3.2 Mining Entailment Relations

In this section we combine our relation-based tuples mined from visual annotations (section 2) with more than 400k textual descriptions included in MISCOCO. We generate a set of entailments \(S_2 = \{r(o_1, o_2) \Rightarrow z\}\), where \(r(o_1, o_2)\) is an element from \(S_1\) and \(z\) is a consequent obtained from textual descriptions. Similarly as in the previous section, we rank the relations in \(S_2\) using a function \(\gamma_2 : S_2 \rightarrow \mathbb{R}\).

We start by generating an exhaustive list of candidate consequents \(z\). We first pre-process the image captions with the part-of-speech tagger and lemmatizer from the Stanford Core NLP toolkit (Manning et al., 2014), and remove stop words. Then we generate a list of \(n\)-length skipgrams in each caption. The set of \(n\)-skipgrams are filtered based on predefined lexical patterns\(^2\), and redundancies are removed\(^3\). Skipgrams, \(z\), are then paired with co-occurring relations, \(r(o_1, o_2)\), removing pairs with the disconnected-from spatial relation (see Figure 1). Pairs are scored with the conditional probability:
\[
\gamma_2(r(o_1, o_2) \Rightarrow z) = \frac{P[z, r(o_1, o_2)]}{P[r(o_1, o_2)]} \tag{2}
\]

The consequent \(z\) can take the form \(q, q(o_1), q(o_2)\), or \(q(o_1, o_2)\), by performing a simple alignment with the arguments in the antecedent. We perform this alignment by mapping the object categories in the antecedent \(r(o_1, o_2)\) to WordNet synsets, and matching any word in \(z\) to any word in the gloss set of the predicate arguments \(o_1\) and \(o_2\). The unmatched words in \(z\) form the relation, whereas matched words form arguments. We produce the form \(q\) if there are no matches, \(q(o_1)\), or \(q(o_2)\) when one argument word matches, and \(q(o_1, o_2)\) when both match. Examples of discovered entailments are in Figure 3.

### 3.3 Generalizing Relations using WordNet

In this section we present an approach to generalize an initial set of relations, \(S\), to objects not found in the original vocabulary \(V\). Using WordNet we construct a superset \(G\) containing all possible parent relations for the relations in \(S\) by replacing their arguments \(o_1, o_2\) by all their possible hypernyms. Our objective is to select a subset \(T\) from \(G\) that contains high quality and diverse generalized relations.

\(^2\langle\text{noun, verb}\rangle, \langle\text{noun, verb}^*, \text{noun}\rangle, \langle\text{noun, preposition}^*, \text{noun}\rangle, \langle\text{noun, verb, preposition}^*, \text{noun}\rangle\)

\(^3\langle\text{noun, verb, noun}\rangle, \langle\text{noun, noun}\rangle, \langle\text{verb, noun}\rangle, \langle\text{noun, verb}\rangle\) are collapsed to \langle\text{noun, verb, noun}\rangle.
holds(chair, cat) on(cat, chair)  
holds(suitcase, cat) in(cat, suitcase)  
holds(person, donut) eating(person)  
on(knife, ... quality. Facts about persons are higher quality, likely because this category is in over 50% of the images in MS-COCO.

Figure 3: Left: correctly identified entailment relations and right: failure cases.

Note that elements in $G$ can be too general and contradict statements in $S$ while others could be correct but add little new knowledge. To balance these concerns, we formulate the selection as an optimization problem by maximizing a fitness function $L$:

$$\max_T L(T), \text{ such that } |T| = k, \text{ and } T \subseteq G,$$

$$L(T) = \lambda \log(1 + \psi(T)) + \sum_{t \in T} \log(1 + \phi(t, S)),$$

where $\psi$ is a coverage term that computes the total number of facts implied through hyponym relationships by the elements in $T$. The second term $\phi$ is a consistency term that measures the compatibility of a generalized relation $t$ with the relations in $S$. We assume that if a relation is missing from $S$, then it is false (this corresponds to a closed world assumption over the domain of $S$). Thus, $\phi$ is the ratio of the scores of relations in $S$ consistent with relation $t$ (i.e. evidence for $t$ based on $S$), and a value that is proportional to the number of missing relations from $S$ (i.e. the amount of counter-evidence). More concretely:

$$\phi(t, S) = \frac{\sum_{s: t \rightarrow s \land s \in S} \gamma(s)}{\mu \cdot (1 + \sum_{s: t \rightarrow s \land s \in S} 1) \cdot d(t, S)},$$

where $\mu$ is a constant and $d$ is the product of the WordNet distances of the synsets involved in $t$ to their nearest synset in $S$. This penalizes relations that are far away from categories in $S$. The optimization defined in Equation 3 is an instance of the submodular k-coverage problem. We use a greedy algorithm that adds elements to $T$ that maximize $L$, which due to the submodular nature of the problem approximates the solution up to a constant factor.

### 4 Experimental Setup

**Object-Object Relations**: We filter out from the initial set of candidate relations the ones that occur less than 20 times. We extract more than 3.1k unique statements (6k including symmetric spatial relations). **Entailment Relations**: We use skipgrams of length 2-6 allowing at most 6 skips, filter candidates such that they occur at least 5 times, and return the top 10 most likely entailments per spatial relation. Overall, 6.3k unique statements are extracted (10k including symmetric relations). **Generalized Relations**: We optimize Equation 4 only for object-object relations because the closed world assumption makes counts for implications sparse. The parameter $\mu$ is set to the average of the scores, $\lambda = 0.05$ and $k = 200$.

### 5 Evaluation

We evaluated the quality of the common sense we derive on Amazon Mechanical Turk. Annotators are presented with possible facts and asked to grade statements on a five point scale. Each fact was evaluated by 10 workers and we normalize their average responses to a scale from 0 to 1. Figure 4 shows plots of quality vs. coverage, where coverage means the top percent of relations sorted by our predicted quality scores.

**Object-Object Relations** As a baseline, 1000 randomly sampled relations have a quality of 0.225. Figure 4a shows our PMI measure ranks many high quality facts at the top, with the top quintile of the ranking being rated above 0.63 in quality. Facts about persons are higher quality, likely because this category is in over 50% of the images in MS-COCO.
Entailment Relations Turkers were instructed to assign the lowest score when they could not understand the consequent of the entailment relation. As a baseline, 1000 randomly sampled implications that meet our patterns have a quality of 0.33. Figure 4b shows that extracting high quality entailment is harder than object-object relations likely because supposition and consequent need to coordinate. Relations involving furniture are rated higher and manual inspection revealed that many relations about furniture imply stative verbs or spatial terms.

Generalized Relations To evaluate generalizations, Figure 4c, we also present users with definitions\(^4\). As a baseline, 200 randomly sampled generalizations from our 3k object-object relations have a quality of 0.53. Generalizations we find are high quality and cover over 400k objects facts not present in MS-COCO. Examples from the 200 we derive include: holds(dining-table, cutlery), holds(bowl, edible fruit) or on(domestic animal, bed).

6 Conclusion
In this work, we use an existing object detection dataset to extract 16k common sense statements about annotated categories. We also show how to generalize using WordNet and induced hundreds of thousands of facts about unseen objects. The information we extracted is visual, large scale and good quality. It has the potential to be useful for both visual recognition and entailment applications.

\(^4\)sometimes rules involve abstract concepts, for example vessel, any object that can be used as a container

7 Acknowledgments
The authors would like to thank Hannaneh Hajishirzi, Yejin Choi, and Luke Zettlemoyer for helpful comments on the draft and the reviewers for useful feedback. We also thank Ani Kembhavi, and Roozbeh Mottaghi for helping to formalize visual common sense and Mechanical Turkers who helped evaluate output.

References


Dekang Lin and Patrick Pantel. 2001. Dirt@ sbt@ discovery of inference rules from text. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pages 323–328. ACM.


