# VisKE: Visual Knowledge Extraction and Question Answering by Visual Verification of Relation Phrases

Fereshteh Sadeghi[†] [*] Santosh K. Divvala[‡,†] Ali Farhadi[†,‡]
[†]University of Washington [‡]The Allen Institute for AI
{fsadeghi,ali}@cs.washington.edu, santoshd@allenai.org

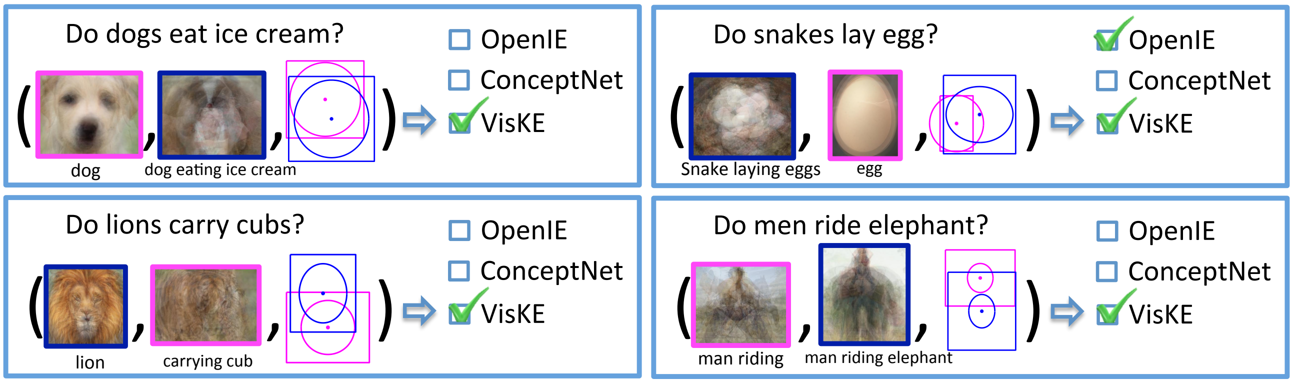

Figure 1. Do dogs eat ice cream? While we humans have no trouble answering this question, existing text-based methods have a tough time. In this paper, we present a novel approach that can visually verify arbitrary relation phrases.

## Abstract

*How can we know whether a statement about our world is valid. For example, given a relationship between a pair of entities e.g., 'eat(horse, hay)', how can we know whether this relationship is true or false in general. Gathering such knowledge about entities and their relationships is one of the fundamental challenges in knowledge extraction. Most previous works on knowledge extraction have focused purely on text-driven reasoning for verifying relation phrases. In this work, we introduce the problem of visual verification of relation phrases and developed a Visual Knowledge Extraction system called VisKE. Given a verb-based relation phrase between common nouns, our approach assess its validity by jointly analyzing over text and images and reasoning about the spatial consistency of the relative configurations of the entities and the relation involved. Our approach involves no explicit human supervision thereby enabling large-scale analysis. Using our approach, we have already verified over 12000 relation phrases. Our approach has been used to not only enrich existing textual knowledge bases by improving their recall, but also augment open-domain question-answer reasoning.*

## 1. Knowledge Extraction & Visual Reasoning

*Do dogs eat ice cream?* If you know the answer to this question, then you either have witnessed dogs eating ice cream or have observed either visual or textual recordings of this phenomenon. Extracting such knowledge has been a long-standing research focus in AI, with a variety of techniques for automatically acquiring information of our world.

Vision is one of the primary modalities for us humans to learn and reason about our world. We gather our everyday basic knowledge such as *horses eat hay* or *butterflies flap wings* by simply observing these phenomenon in our real world. Yet when extracting knowledge for building intelligent systems, most previous research has focused on reasoning primarily using language and text.

Why such a disconnect? This disparity has mainly stemmed from the fact that we have had easier access to copious amounts of text data on the web along with well-performing feature representations and efficient unsupervised learning methods for text. However this does not hold true any more. Thanks to the proliferation of camera phones and the popularity of photo-sharing websites, recent years have witnessed a deluge of images on the web. Coupled with the growing success of text-based image search engines and the recent progress in weakly-supervised object localization methods, we believe the time is ripe now for extracting knowledge by reasoning with images.

[*] This work was done while the author was an intern at the Allen Institute for AI.

**Problem Overview:** The key component of any knowledge extraction system involves verifying the validity of a piece of gathered information before adding it to a knowledge base. The most typical format of the information being considered is in the form of a *relationship* between a pair of *mentions* e.g., *eat(horse, hay)*, *flutter(butterfly, wings)*, etc.

The primary focus of our work is to estimate the confidence of such mentions-relation predicates by reasoning with images. We focus our attention to verb-based relations between common nouns. The input to our system is a relation predicate e.g., 'eat(horse, hay)' and the output is a confidence value denoting its validity. In order to correctly validate a relation, we need to reason about the underlying entities while grounding them in the relation being considered. Here, we present a novel verification approach that reasons about the entities in the context of the relation being considered using webly-supervised models for estimating the spatial consistency of their relative configurations.

The attractive feature of our proposed framework is that both our model learning as well as inference steps are performed using no explicit human supervision. Most previous research on analyzing objects and their relationships in computer vision have assumed a supervised setting i.e., images along with some annotations of the objects and actions involved are available at training. This limitation has prevented these methods to scale to a large number of objects and relations. Our proposed approach overcomes this limitation by carefully leveraging unlabeled images found on the web, thereby enabling image-based reasoning for knowledge extraction.

In summary, our key contributions are: (i) We introduce the problem of visual verification of relation phrases for the task of knowledge extraction. (ii) We present an unsupervised approach for verifying relationships by analyzing the spatial consistency of the relative configurations of the entities and the relation involved. (iii) We empirically demonstrate the utility of our approach on a large relation phrase dataset and analyze the relative contributions of the different system components. (iv) To date, we have verified over 12000 relation phrases and doubled the size of the ConceptNet knowledge base [34] at a precision of 0.85. (v) We released our data and system for enabling future research and applications in this direction. (We invite the interested reader to verify a relation phrase of their choice using our online system.)

## 2. Related Work

The task of verifying relation phrases has received extensive attention in the field of information extraction. Phenomenal progress has been achieved using a variety of methods [1, 2, 4, 5, 11, 29, 36]. The core idea behind these methods involves analyzing the frequency of occurrence of a given relation predicate in large text corpora [2, 11]. While frequency of occurrence in text is a reasonable indicator for the validity of a relation, it is not completely fool-proof. Many high frequency relations occur in text but are not true in the real world e.g., 'pierce(pelican, breast)'. Conversely many relations occur in text with low frequency but are true in the real world e.g., 'eat(chimpanzee, icecream)'. This anomaly springs from the fact that we humans often fail to explicitly state (in text) several obvious pieces of knowledge [18, 37] and therefore text-based methods can miss many basic relationships. Nonetheless these phenomenon are captured in the photos that we take in our daily life. Therefore by reasoning with images, we can leverage complementary cues that are hard to gather purely from text.

However, the task of relation verification has not yet received much attention in computer vision. Most previous research in this area has primarily focused on the tasks of image classification [8], scene recognition [10, 27, 31, 38] and object detection [12, 16, 17, 30] that form the fundamental building blocks for higher order visual reasoning systems. Subsequent research has leveraged the success in the classification and detection tasks to gather structured visual knowledge about objects, scenes and other concepts on an Internet scale [6, 9]. Also, in [41] the problem of learning common sense knowledge from clip art images was studied.

Recent years have also witnessed a surge in reasoning about human-object relationships [19, 39] as well as more general object-object relationships [16, 23, 24, 25] and object-attribute relationships [6, 15, 40]. However, almost all works have studied this problem in the supervised setting i.e., images along with some form of annotations for the objects and the actions involved are assumed to be provided during training.

Quite related to our work is the work of [6], wherein the goal was to extract common-sense relationships between objects in an unsupervised setting. By analyzing the co-detection pattern between a pair of objects, the relationship between them was determined. Their focus was on two types of relationships: 'part of' and 'similar to'. In this work, we attempt to generalize their goal by learning and extracting more general relationships. We show it is possible to learn arbitrary relationships (such as 'eat', 'ride', 'jump', etc.,) by reasoning about the objects in the context of the relation connecting them. We have used our method to learn over 1100 relation types. Our work is complementary to the work of [40], where the utility of a knowledge base of relationships for performing diverse set of visual inference tasks was demonstrated. The knowledge gathered by our work can help enrich their underlying knowledge base, thereby facilitating more advanced inference tasks.

## 3. Visual Verification

Our key assumption in visual verification is that true relation phrases are those that happen in our real world and therefore there should exist enough visual recordings (images, videos) of them online. We consider visual verifica-

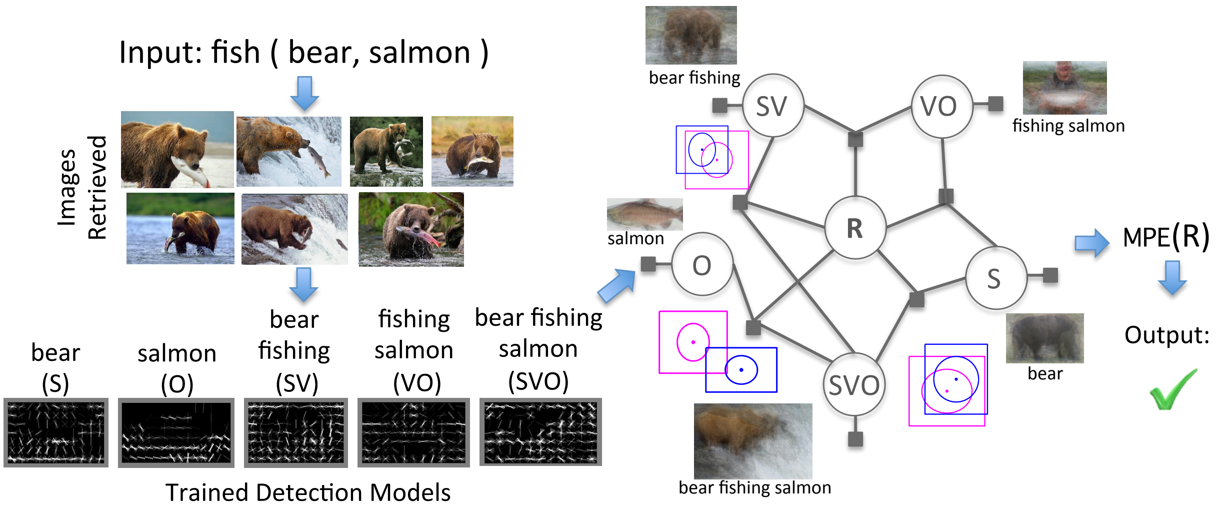

Figure 2. Approach Overview. Given a relation predicate, such as fish(bear,salmon) VisKE formulates visual verification as the problem of estimating the most probable explanation (MPE) by searching for visual consistencies among the patterns of subject, object and the action being involved.

tion of a relation phrase as the problem of searching for meaningful and consistent patterns in the visual recordings of the relation. But what are these patterns of consistencies for relation phrases?

For example, given a relation predicate, such as *fish(bear,salmon)* (read as Do bears fish salmon?) in Figure 2, we observe an *open-mouthed bear* attempting to catch hold of a *leaping salmon*, with the salmon appearing *in front* of the bear. This observation leads us to the following valuable insights about the subject and object involved in a relationship.

First, the appearance of the subject as well as the object may change during the relationship. In this example, we see that the appearance of the bear while fishing salmon is different from a canonical appearance of a bear (i.e., open-mouthed). Similarly the appearance of a salmon when being fished is different from its canonical appearance (i.e., leaping). Therefore to find the occurrence of the subject and the object pattern (i.e., bear or salmon) in the images, it is important to search not only for their canonical patterns but also for their patterns under the relationship in consideration (i.e., 'bear fishing' and 'fishing salmon'). This change in the appearance due to interaction is aligned with the idea of visual phrases [32].

Second, the spatial locations of the subject and the object should be in a consistent behavior for the relationship to hold. In this example, we see that the salmon consistently appear in a specific location with respect to the bear for the relationship of fishing to be valid. Therefore to validate a relationship between a subject and an object, we need to check for the spatial consistency between the subject (bear) and the object (salmon) patterns and also between their modified versions (i.e., 'bear fishing', 'fishing salmon'). In the following, we present our formulation that generalizes these intuitions.

We refer to a relation as $\mathcal{V}(\mathcal{S},\mathcal{O})$, where $\mathcal{V}$ denotes the verb or the action connecting the subject $\mathcal{S}$ and the object $\mathcal{O}$). Based on our observations, participation in a relationship changes the appearance of the participating entities in different ways. For a relation $\mathcal{V}(\mathcal{S},\mathcal{O})$, we envision the following possibilities of meaningful patterns ($\mathcal{R}$): First, a relation ($\mathcal{V}$) might form a visual phrase and change the appearance of the subject ($\mathcal{S}$) i.e., ($\mathcal{SVO}, \mathcal{SV}$); Second, the relation might affect the object ($\mathcal{O}$) i.e., ($\mathcal{S}, \mathcal{VO}$); Third, the relation might form a visual phrase and change the appearance of both subject and object i.e., ($\mathcal{VO}, \mathcal{SV}$); Fourth, the relation might impose specific visual characteristics on the subject but the object is not affected i.e., ($\mathcal{SVO}, \mathcal{O}$); and Fifth, the relation might impose specific visual characteristics on the object but the appearance of the subject remains intact i.e., ($\mathcal{SVO}, \mathcal{S}$). We ignored the $\mathcal{V}$, $\mathcal{SO}$ variables as in isolation they are highly visually ambiguous. We enforced the participation of all the three $\mathcal{S}$, $\mathcal{V}$, $\mathcal{O}$ entities in patterns and therefore avoide patterns like ($\mathcal{S}, \mathcal{SV}$) as it does not involve the $\mathcal{O}$.

Searching for consistencies among the above patterns require detectors for each of the elements of relations i.e., the subject ($\mathcal{S}$), the object ($\mathcal{O}$), the subject-verb combination ($\mathcal{SV}$), the verb-object combination ($\mathcal{VO}$), and the subject-verb-object combination ($\mathcal{SVO}$).

Assuming we have access to these individual detection models (explained later), we formulate visual verification as the problem of estimating the most probable explanation (MPE) of the multinomial distribution that governs $\mathcal{R}$. We factorize the marginalization of the joint distribution of $\mathcal{R}$ and the relation elements using a factor graph (depicted in Figure 2):

$$P(\mathcal{R},\mathcal{S},\mathcal{O},\mathcal{SV},\mathcal{VO},\mathcal{SVO}) \propto \prod_{x\in\{\mathcal{O},\mathcal{S},\mathcal{SV}\}} \Phi(\mathcal{R},\mathcal{SVO},x) * \prod_{y\in\{\mathcal{SV},\mathcal{S}\}} \Phi(\mathcal{R},\mathcal{VO},y) * \prod_{z\in\{\mathcal{S},\mathcal{O},\mathcal{SV},\mathcal{VO},\mathcal{SVO}\}} \Psi(z), \quad (1)$$

where $\mathcal{R}$ corresponds to the relation type and has a multinomial distribution over the patterns of consistency, the rest of the nodes correspond to relation element detectors. The potential function $\Phi$ provides the maximum likelihood estimates of each relation type. More specifically,

$$\Phi^i(\mathcal{R},x,y) = \begin{cases} \max_\theta \mathcal{L}(x,y,\bar{I};\theta) & \mathcal{R}\equiv i \\ 1 & \text{otherwise} \end{cases} \quad (2)$$

where $\bar{I}$ is the set of images collected for a relation phrase, and $\mathcal{L}(x,y,\bar{I},\theta)$ is the log likelihood of the parametric relations between detections of $x$ and $y$ on image set $\bar{I}$ parameterized by $\theta$. For the parametric models we use Gaussians. The $\Psi(x)$ is the unary factor representing the maximum log likelihood estimates of predictions of detector $x$. Referring back to the example of bear fishing salmon, our factor graph checks for at least one of the five patterns to hold i.e., either 'bear fishing' and 'fishing salmon', or 'bear' and 'fishing salmon', or 'bear fishing salmon' and 'bear', 'bear fishing salmon' and 'salmon' or 'bear fishing salmon' and 'bear fishing' should have a highly consistent pattern for this relationship to be valid.

The features to estimate the maximum likelihood estimate $\mathcal{L}(x,y,\bar{I};\theta)$ should capture the spatial relations between the predictions of detectors $x$ and $y$. Towards this end, we use the following feature representation (See Figure 3): $\{dx,dy,ov,ov_1,ov_2,h_1,w_1,h_2,w_2,a_1,a_2\}$, where $dx,dy$ correspond to the translation between detections, $ov$ is the intersection over the union of the two detection boxes, $ov_1$ is the ratio of intersection over the area of the bounding box $x$, $ov_2$ is the ratio of the intersection over the area of bounding box $y$, $h_1,w_1$ are the dimensions of the bounding box $x$, $h_2,w_2$ are the dimensions of the bounding box $y$ and $a_1,a_2$ are the $x$ and $y$ bounding box areas. For unary potentials we use a 4 dimensional representation that encodes $\{h,w,x,y\}$, where $h,w$ are the height and width of the bounding box and $x,y$ are its (mid-point) coordinates. Under this model, visual verification is the problem of MPE in our factor graph [28].

**Implementation Details:** We use the publicly-available implementation of [9] for learning our $\mathcal{S}$, $\mathcal{O}$, $\mathcal{SV}$, $\mathcal{VO}$, $\mathcal{SVO}$ detectors (without parts). For each detector, a mixture of components Deformable Part Model (DPM) [16] is trained using retrieved images from the web and the noisy components are pruned in a separate validation step. [1] As our individual detectors (i.e., $\mathcal{SV}$, $\mathcal{SVO}$, $\mathcal{S}$, $\mathcal{O}$, $\mathcal{VO}$) are mixture models, each of our factors (e.g., $(\mathcal{SV},\mathcal{SVO})$) incorporate these mixtures.

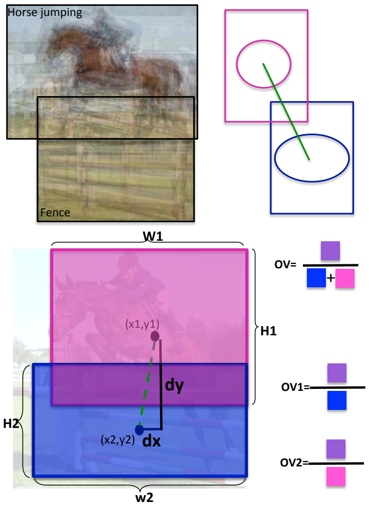

Figure 3. Our feature representation used for estimating the parametric relation between two detection elements. This figure shows the feature computed between 'Horse jumping' ($\mathcal{SV}$) and 'Fence' ($\mathcal{O}$). The ellipses on top show the distribution of spatial position of 'Horse jumping'(in pink) with respect to 'Fence'(in blue) as learned by our method. There is high horizontal variation in the position of 'Fence' compared to its vertical variation while the spatial position of Horse jumping is above 'Fence' and has small horizontal and vertical variation.

## 4. Relation Phrase Dataset

To the best of our knowledge, there exists no gold-standard dataset of relation phrases in the knowledge extraction community. Different extraction systems use different strategies for extracting relational phrases [1, 2, 5]. To avoid biasing our proposed approach to the idiosyncracies of any one of these methods, we have put together a generic dataset of relation phrases.

Our relation phrases were gathered using the Google Books Ngram (English 2012) corpus [26]. This corpus contains parts-of-speech tagged Ngrams along with their frequency of occurrence in books. To extract relations of 'verb(subject, object)' format, we considered all Ngrams that have a <noun, verb, noun> pattern (ignoring any other words in between). For the purpose of our experiments, we focused our attention to the relations involving animals. This resulted in a list of 6093 relations covering 45 different subjects, 1158 different verbs (actions) and 1839 different objects. To avoid biasing this list to contain only true relations, we generated new relations by randomly permuting the subjects, verbs and objects together yielding an additional 6500 relations (resulted in a total of 5776 pairs of $\mathcal{SV}$ and 5186 pairs of $\mathcal{VO}$ in our dataset). We refer to these relations as the 'Permute' set and the former as the 'Base' set. Some of the sample relations can be seen in Figures 1 2 5 8.

[1] Using our current unoptimized linux-based Matlab implementation on a Intel Xeon E5 CPU, the entire run-time per relation is around 30mins.

|  | Base Set | Permute Set | Combined Set |
|---|---|---|---|
| Visual Phrase [32] | 49.67 | 14.12 | 42.49 |
| Co-detection Model | 49.24 | 14.65 | 43.14 |
| Google Ngram Model [1] | 46.17 | NA | NA |
| Language Model [22] | 56.20 | **22.68** | 50.23 |
| VisKE | **62.11** | 20.93 | **54.67** |

Table 1. Results (M.A.P.) on the Relation Phrase Dataset. While the language model achieves a higher accuracy on the 'Permute' set, VisKE gets the best result on the 'Base' set and the 'Combine' set.

For evaluating the performance of our method as well as to compare different baselines, each relation phrase was annotated with its ground-truth validity. The validity was primarily decided based on whether a relations refers to a phenomenon that commonly happens in our real-world. Out of the total 12593 relations, 2536 statements were annotated as true and the rest as false[2].

## 5. Experimental Results & Analysis

We analyzed the efficacy of our approach by conducting experiments on the relation phrase dataset. The input to our approach is a relation phrase e.g., 'eat(horse, hay)' and the output is a confidence score denoting its validity (i.e., larger value indicates greater confidence in being true). We use these scores along with their corresponding ground-truth labels to compute a precision-recall curve and use the area under the precision-recall curve (Average Precision, A.P.) metric [12] as a principal quantitative measure. We computed the A.P. independently for each subject and then averaged the A.P. across all subjects (Mean A.P., M.A.P.). Table 1 summarizes the key results obtained using our approach. We separately evaluated the A.P over the 'Base' set and the 'Permute' set to analyze the impact in the different scenarios. Our proposed approach achieves an M.A.P. of 62.11% on the 'Base' set, and 20.93% on the 'Permute' set, indicating the difficulty of the latter compared to the former. Validating 12593 relations involved training over 26739 detectors and processing around 9 million images. Our experiments were run on a computer cluster. We also compared our results to the following baseline models to analyze different aspects of our system.

*Co-detection model:* We first compared against a simple approach that trains separate detection models for the entities and the relation involved (using the web images) and then analyzes the pattern of their co-detections. For example, in the case of 'eat(horse, hay)', separate detection models were trained for horse and hay as well as a detector for the relation eat (using their corresponding web images) and then reasoned about the pattern of their co-detections on 'horse eating hay' images. This approach is most similar to that of [6]. As seen in Table. 1 (row2), this approach fails to perform well as it considers each constituent of the relation phrase independently and thereby fails to account for the changes in appearance of the entities when involved together in an action [32]. For example, in case of the horse eats hay example, the appearance of the 'eating horse' is different from that of a canonical horse.

*Visual Phrase model:* We next compared our approach to the visual phrase model of [32], where a single $\mathcal{SVO}$ detector is trained and its performance on predicting its unseen samples is evaluated. As seen in Table. 1 (row1), this model fares poorly. We found it to classify several incorrect relations to be true as it does not validate the pattern of its constituent entities. For example, in case of 'horse read book', the retrieved images contain cover pages of books (about horses) all having a picture of horse[3]. Validating a detector on these images would lead to falsely claiming this relation to be true as it just analyzes the consistency of the overall pattern without reasoning about the action of horse reading.

*Language model:* Of course our task of visually verifying relation phrases has been motivated from the domain of text-driven knowledge extraction. It is therefore interesting to compare our method to contemporary text-driven methods for verifying relation phrases. We evaluated the performance of two popular industrial-sized language models (Bing, Google). The method of [22] estimates the real-world plausibility of any relation phrase using a sophisticated statistical language model learned from a large text corpora, while the method of [1] estimates the probabilities using a language model trained on the GoogleNgram corpus. As seen in Table. 1, although the language model of [22] outperforms the co-detection and phrasal baselines, it does not perform as well as our proposed approach.

To analyze performance per subject, in Figure. 4, we display the individual relative differences in A.P. Out of the 45 subjects, our approach does better on 27 of them, while the language model of [22] does better on 14. For subjects like 'pelican', 'lizard' etc., our approach does better, while for subjects like 'pig', 'monkey', language model does better. We hypothesize this to the fact that classes like monkey are more common than classes like pelican and therefore the language model has more data for these classes. This difference in performance between our model and the language model hints at the complementarity of the vision and language methods. To validate this hypothesis, we ran a separate evaluation (on the Permute set) by linearly combining the confidences produced by these two methods. This combined model indeed produced a higher M.A.P. of 24.25%[4], ascertaining the fact that reasoning with images offers cues complementary to text for relation phrase verification. As the number of relation phrases per subject in our dataset is

[2] We have released our list of relational relations along with their annotations in our project website (http://viske.allenai.org/).

[3] This phenomenon happens as image-search engines predominantly rely on auxiliary text (around the image) in the documents for retrieving images.

[4] The performance on the Base set did not improve. We hypothesize this due to our simple linear combination strategy. Given the different score calibrations of the visual and language model, it is a challenge to combine them meaningfully in an unsupervised setting.

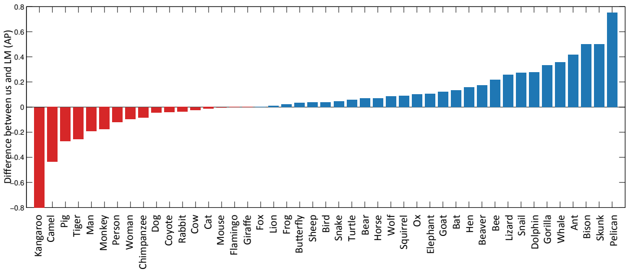

Figure 4. Performance improvement by VisKE over the language model [22] (x-axis: 45 subjects, y-axis: difference in A.P.). Blue indicates our approach does better than [22] (27/45 classes), while red indicates vice versa (14/45 classes).

| Model | M.A.P. |
|---|---|
| VisKE (All Factors) | 62.11 |
| Without $\Phi(\mathcal{R}, \mathcal{VO}, \mathcal{SV})$ | 60.41 |
| Without $\Phi(\mathcal{R}, \mathcal{VO}, \mathcal{S})$ | 61.16 |
| Without $\Phi(\mathcal{R}, \mathcal{SVO}, \mathcal{S})$ | 60.40 |
| Without $\Phi(\mathcal{R}, \mathcal{SVO}, \mathcal{O})$ | 59.55 |
| Without $\Phi(\mathcal{R}, \mathcal{SVO}, \mathcal{SV})$ | 59.55 |
| Without binary terms | 60.61 |
| Without unary terms | 58.52 |
| CRF | 58.01 |

Table 2. Ablation analysis: while each of the factors help improving the overall performance, removing any of them does not drastically hurt its performance, indicating the robustness of our overall model. Removing either of the binary or unary terms hurts the performance. Using a CRF model results in poorer performance.

imbalanced, we also evaluated the performance over the entire set of relation phrases (across all subjects) on the 'Base' set. This yielded an A.P. of 44.6% for our method, while the LM [17] obtained 40.2%.

**Ablation Study:** To understand which factors within our model are critical towards the final performance, we analyzed results by running experiments with the different factors turned off/on. As displayed in Table. 2, while each of the factors helps in improving the overall performance, removing any one of them does not drastically hurt the performance, indicating the robustness of our overall model. Also, as observed in Table 2, both the unary and the binary factors contribute towards the final performance. We also ran a separate experiment where we used a simple CRF based pooling strategy to combine the responses of the different pattern relationships i.e., $(\mathcal{SVO}, \mathcal{SV})$, $(\mathcal{VO}, \mathcal{SV})$, etc., which resulted in poorer performance.

What are the sources of errors that prevent our model in correctly verifying some of the relationships? We found a couple of issues. First, our method is dependent on web image search engines to gather the relevant set of images. For some relations, e.g. make(fox,den), the retrieved images are not relevant, while for some other relations, e.g. shed(cow, horn), the images are misleading. Second, our method uses the webly-supervised approach of [9] for training the detectors, which sometimes fails either when the variance within the set of retrieved images is large, e.g. eat(horse, fruit), or if the relation involves complex visual characteristics, e.g. drink(hedgehog, milk). Finally, the inherent spatial relationships in case of certain relation phrases is complex, e.g. cross(horse, road). Verifying such relations require deeper understanding of spatial relations. Future work could explore leveraging (textual) prepositions to better understand complex spatial relationships.

## 5.1. Application: Enriching Knowledge Bases

Current knowledge bases such as WordNet, Cyc, ConceptNet, etc., seldom extract common-sense knowledge directly from text as the results tend to be unreliable and need to be verified by human curators. Such a manual process is both labor intensive and time consuming. A nice feature of our method is that it offers complementary and orthogonal source of evidence that helps in discovering highly confident facts from amongst the pool of all facts extracted from text. This feature helps us towards automatically improving the recall of knowledge bases. We demonstrate this feature on the popular ConceptNet knowledge base.

ConceptNet is a semantic network containing common-sense knowledge collected from volunteers on the Internet since 2000 [34]. This knowledge is represented as a directed graph whose nodes are concepts and edges are assertions of common sense relations about these concepts. The set of possible relations is chosen to capture common informative patterns, such as 'IsA', 'PartOf', 'HasA', 'MemberOf', 'CapableOf', etc.

In our analysis, we measured the number of relation predicates our visual approach could add to this resource. More specifically, for each of the 45 subjects in the relation phrase dataset, we measured the precision of correct relationships (that are unavailable in ConceptNet) added at different levels of recall. While some of the relationships (e.g., 'IsA(horse, animal)', 'PartOf(wheel, car)', 'HasA(horse, leg)') are easier to acquire and have been previously explored in computer vision [35, 25, 6], more complex *action-centric* relationships (e.g., 'CapableOf(horse, jump fence)') have not received much attention. In fact, to our surprise, across the 45 concepts there were only 300 'CapableOf' relation facts within ConceptNet[5]. In our analysis, we primarily focused on increasing the number of facts pertaining to this relationship.

Figure. 6 summarizes the key results of our analysis. It displays the number of relations added at various precision levels for four (of the 45) subjects. For example, in case of 'cat', we have added 10 new relations to ConceptNet at a precision of 0.8. Some of the newly added relations are (i.e., unavailable in ConceptNet): 'CapableOf(cat, lick kitten)', 'CapableOf(cat, carry mouse)', 'CapableOf(cat, sit basket)', 'CapableOf(cat, chase bird)', and 'CapableOf(cat, lay cushion)'. On average, at a precision of .85, we doubled the size of the 'CapableOf' relations in ConceptNet (related

[5] ConceptNet is an everchanging repository. The results here correspond to the latest version downloaded on September 26 2014.

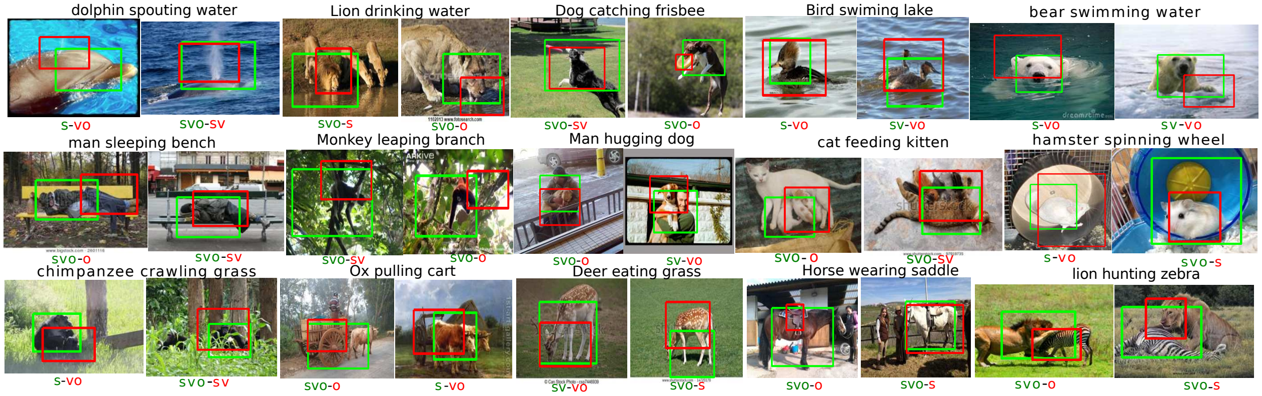

Figure 5. Examples of detected entities in a few of the relation phrases verified by VisKE. Entities are color coded within pattern.

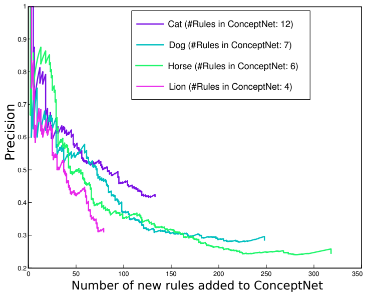

Figure 6. Enriching ConceptNet Knowledge Base: Figure shows the number of correct relationships added by VisKE at different levels of recall. For example, in case of 'cat', we have added 10 new relations to ConceptNet at a precision of 0.8.

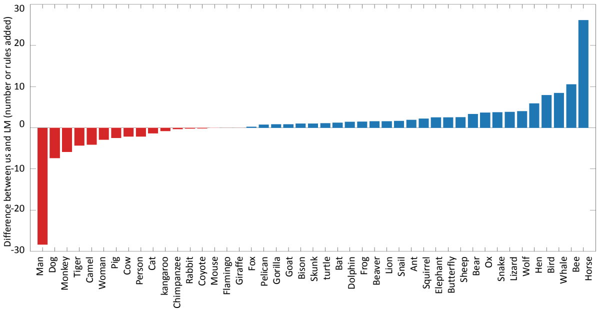

Figure 7. Relative difference of the number of added relations to ConceptNet using VisKE over the language model [22] (x-axis: 45 subjects, y-axis: difference in number of relations). Blue indicates that VisKE adds more relations than [22] (28/45 classes), while red indicates vice versa (14/45 classes).

to our subjects). In Figure. 7, we compare the number of relations added by our approach with respect to the language model [22]. Out of the 45 subjects, our approach adds more relations on 28 of them, while the language model [22] does better on 14. This visualization again reveals the the complementarity in performance between our model and the language model.

*OpenIE:* We conducted a similar analysis on OpenIE [2, 13], a popular web-scale information extraction system that reads and extracts meaningful information from arbitrary web text. We selected the relations corresponding to the 45 subjects of interest within their extractions [13] and ran our approach to compute the confidences. Table. 3 summarizes our results. Our approach helps in improving the extractions obtained by OpenIE as well.

| Model | M.A.P. |
|---|---|
| OpenIE [13] | 73.03 |
| Co-detection Model | 76.65 |
| Visual Phrase [32] | 78.45 |
| Language Model [22] | 83.65 |
| VisKE | 85.80 |

Table 3. Results on the OpenIE dataset.

*Towards Higher-order Reasoning:* An interesting feature enabled by our approach is reasoning about higher-order relationships. Figure. 9 shows the relation graph learned by our approach based on the learned spatial models for the different relation phrases involving the relationship of 'pushing'. These higher-order relations were estimated by computing the cosine similarity of the different pairs of relations based on the maximum similarity of their corresponding patterns (i.e., $(\mathcal{SVO}, \mathcal{SV})$, $(\mathcal{VO}, \mathcal{SV})$, etc.,) in the factor graph. The cosine similarity is computed using the feature representation as explained in section 3. The relation graph reveals that the action of 'man pushing box' is very similar to 'woman pushing cart', but not to 'person pushing bicycle'. Such reasoning about similarities and dissimilarities between relationships (in the context of the entities involved) is valuable for several tasks in vision and language.

## 5.2. Application: Question Answer Reasoning

Question-Answering is an important AI problem where the goal is to answer questions by querying large knowledge bases [3, 14]. The visual knowledge that we have gathered using our approach can help improve the reasoning within question-answering systems. For example, as shown in Figure. 8, it could be possible for users to explore and discover a variety of interesting facts about concepts and their relationships, such as:

*What do dogs eat?* Given a query of the form

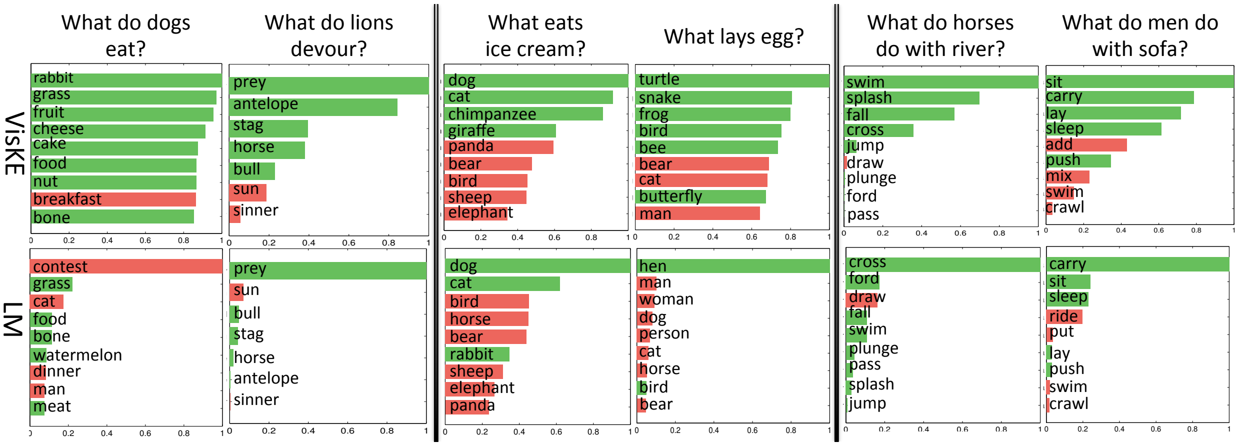

Figure 8. Diverse Question Answering: VisKE is capable of answering diverse questions about subjects, objects or actions. In comparison to the Language model [22], VisKE obtains richer and more precise answers to the questions.

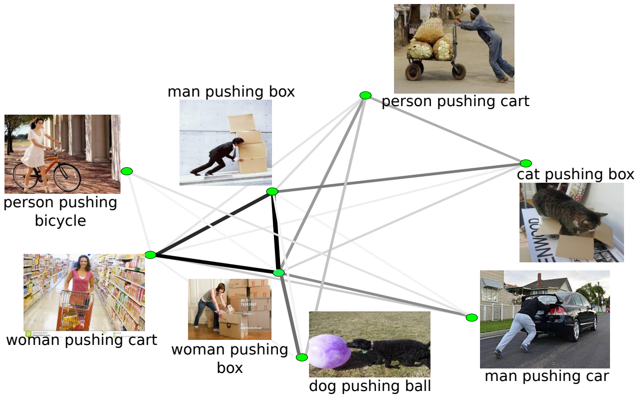

Figure 9. Higher-order reasoning of relationships: This relation graph reveals to us that the action of ‘man pushing box’ is very similar to ‘woman pushing cart’, but not to ‘person pushing bicycle’. The edge thickness shows the strength of a relation. Darker edge indicates higher similarity.

‘verb(subject,?)’, we can use our approach to retrieve the top set of objects relevant to it. For e.g., our approach reveals that dogs typically eat rabbit, grass, fruit, etc. In contrast, the LM has produced high probability for ‘contest’ as it is confused by ‘hot dog eating contest’.

*What lays eggs?* Given a query of the form ‘verb(?, object)’, we can use our approach to retrieve the top set of subjects relevant to it. For e.g., our approach reveals that the most probable animals that can lay eggs are turtle, snake, etc.

*What do horses do with river?* Given a query of the form ‘?(subject, object)’, our approach can retrieve the top set of relations between the subject and object. For e.g., our approach reveals that the most probable relationship between butterfly and wings are flutter and flap, and between man and sofa are sit, carry, sleep, push, etc.

**Towards Answering Elementary-level Science Questions:** Apart from answering generic questions, our approach can be useful for answering more specific questions such as those related to elementary-level general science [7, 21, 33, 20]. To demonstrate this, we ran a preliminary experiment using our approach on a small subset of the NewYork Regents’ 4th grade science exam [7] that were visually relevant. Given a question such as ‘What part of a plant produces seeds? (a) Flower, (b) Leaves, (c) Stem, (d) Roots’, it is decomposed into its constituent relations[6] i.e., ‘produce(flower, seeds)’, ‘produce(leaves, seeds)’, ‘produce(stem, seeds)’, ‘produce(roots, seeds)’. Our approach validates each of the relations and outputs a confidence value for them. This confidence is used to pick the right answer. Our approach achieved an accuracy of 85.7% (compared to 71.4% by a text-driven reasoning approach [7]) highlighting the benefit of our visual reasoning based question answering approach.

## 6. Conclusion

Relation verification constitutes a fundamental component within any knowledge extraction system. In this paper, we have highlighted the importance of visual reasoning for relation phrase verification. We have presented a novel approach for visual verification that reasons about the entities in the context of the relation being considered, by estimating the spatial consistency of their relative configurations using webly-supervised models. Using our approach, we have demonstrated impressive results on a large relation phrase dataset and also highlighted the complementarity of the cues provided by our approach in comparison to existing linguistic models. Further we have also demonstrated the utility of our approach in enriching existing knowledge bases and visual question answering.

**Acknowledgments:** This work was supported by ONR N00014-13-1-0720, NSF IIS-1218683, NSF IIS-IIS-1338054, and Allen Distinguished Investigator Award.

[6] The relations for the questions were manually created. Automatic relations generation is a challenging research problem [7].

## References

[1] A. Akbik and T. Michael. The weltmodell: A data-driven commonsense knowledge base. In *LREC*, 2014. 2, 4, 5

[2] M. Banko et al. Open information extraction from the web. In *IJCAI*, 2007. 2, 4, 7

[3] J. Berant, A. Chou, R. Frostig, and P. Liang. Semantic parsing on freebase from question-answer pairs. In *EMNLP*, 2013. 7

[4] A. Carlson, J. Betteridge, E. R. Hruschka, Jr., and T. M. Mitchell. Coupling semi-supervised learning of categories and relations. In *NAACL*, 2009. 2

[5] A. Carlson et al. Toward an architecture for never-ending language learning. In *AAAI*, 2010. 2, 4

[6] X. Chen, A. Shrivastava, and A. Gupta. NEIL: Extracting visual knowledge from web data. In *ICCV*, 2013. 2, 5, 6

[7] P. Clark, P. Harrison, and N. Balasubramanian. A study of the knowledge base requirements for passing an elementary science test. In *AKBC*, 2013. 8

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 2

[9] S. Divvala, A. Farhadi, and C. Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *CVPR*, 2014. 2, 4, 6

[10] C. Doersch, A. Gupta, and A. A. Efros. Mid-level visual element discovery as discriminative mode seeking. In *NIPS*. 2013. 2

[11] D. Downey, O. Etzioni, and S. Soderland. A probabilistic model of redundancy in information extraction. In *IJCAI*, 2005. 2

[12] M. Everingham et al. The PASCAL Visual Object Classes (VOC) challenge. In *IJCV*, 2010. 2, 5

[13] A. Fader, S. Soderland, and O. Etzioni. Identifying relations for open information extraction. In *EMNLP*, 2011. 7

[14] A. Fader, L. Zettlemoyer, and O. Etzioni. Open question answering over curated and extracted knowledge bases. In *KDD*, 2014. 7

[15] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009. 2

[16] P. Felzenszwalb et al. Object detection with discriminatively trained part based models. *PAMI*, 2010. 2, 4

[17] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 2

[18] P. Grice. Logic and conversation. In *Speech Acts*, 1975. 2

[19] A. Gupta, A. Kembhavi, and L. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. In *PAMI*, 2009. 2

[20] B. Hixon, P. Clark, and H. Hajishirzi. Learning knowledge graphs for question answering through conversational dialog. In *NAACL*, 2015. 8

[21] M. J. Hosseini, H. Hajishirzi, O. Etzioni, and N. Kushman. Learning to solve arithmetic word problems with verb categorization. In *EMNLP*, 2014. 8

[22] J. Huang et al. Exploring web scale language models for search query processing. In *WWW*, 2010. 5, 6, 7, 8

[23] H. Izadinia, F. Sadeghi, and A. Farhadi. Incorporating scene context and object layout into appearance modeling. In *CVPR*, 2014. 2

[24] T. Lan et al. From subcategories to visual composites: A multi-level framework for object detection. In *ICCV*, 2013. 2

[25] T. Malisiewicz and A. A. Efros. Beyond categories: The visual memex model for reasoning about object relationships. In *NIPS*, 2009. 2, 6

[26] J.-B. Michel et al. Quantitative analysis of culture using millions of digitized books. In *Science*, 2010. 4

[27] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *ICCV*, 2011. 2

[28] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., 1988. 4

[29] A. Ritter, Mausam, and O. Etzioni. A latent dirichlet allocation method for selectional preferences. In *ACL*, 2010. 2

[30] O. Russakovsky et al. Detecting avocados to zucchinis: what have we done, and where are we going? In *ICCV*, 2013. 2

[31] F. Sadeghi and M. F. Tappen. Latent pyramidal regions for recognizing scenes. In *ECCV*, 2012. 2

[32] M. Sadeghi and A. Farhadi. Recognition using visual phrases. In *CVPR*, 2011. 3, 5, 7

[33] M. J. Seo, H. Hajishirzi, A. Farhadi, and O. Etzioni. Diagram understanding in geometry questions. In *AAAI*, 2014. 8

[34] R. Speer and C. Havasi. ConceptNet 5: A large semantic network for relational knowledge. In *http://conceptnet5.media.mit.edu*, 2013. 2, 6

[35] E. Sudderth, A. Torralba, W. T. Freeman, and A. Wilsky. Learning hierarchical models of scenes, objects, and parts. In *Proc. ICCV*, 2005. 6

[36] P. Talukdar, D. T. Wijaya, and T. M. Mitchell. Acquiring temporal constraints between relations. In *CIKM*, 2012. 2

[37] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *SIGCHI*, 2004. 2

[38] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 2

[39] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010. 2

[40] Y. Zhu, A. Fathi, and L. Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *ECCV*, 2014. 2

[41] C. Zitnick, R. Vedantam, and D. Parikh. Adopting abstract images for semantic scene understanding. *PAMI*, 2014. 2