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

Fereshteh Sadeghi[†], Santosh K. Divvala[‡,†], Ali Farhadi[†,‡]
[†]Department of Computer Science & Engineering, University of Washington. [‡]The Allen Institute for AI.

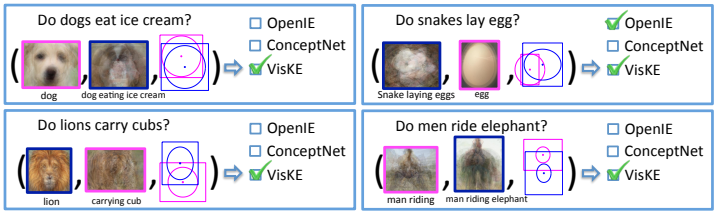

Figure 1: Do dogs eat ice cream? While we humans have no trouble answering this question, existing text-based methods have a tough time. In this paper, we present a novel approach that can visually verify arbitrary relation phrases.

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. 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. 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 develop a novel method that can visually verify the validity of a *relationship* between a pair of *mentions* (e.g., *eat(horse, hay)*, *flutter(butterfly, wings)*). Given such a verb-based relation phrase between common nouns, our approach assess its validity by jointly analyzing over text and natural images and reasoning about the spatial consistency of the relative configurations of the entities and the relations. 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. This allows our system to scale up to a large number of relations. 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.

Visual Verification: The primary focus of our work is to estimate the confidence of mentions-relation predicates by reasoning with natural images. We focus our attention to verb-based relations between common nouns. The input to our system is a mentions-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. In this paper, 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. Searching for consistencies among the 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, 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 Φ provides maximum likelihood estimates of each relation type. The $\Psi(x)$ is the unary factor representing the maximum log likelihood estimates of predictions of detector x.

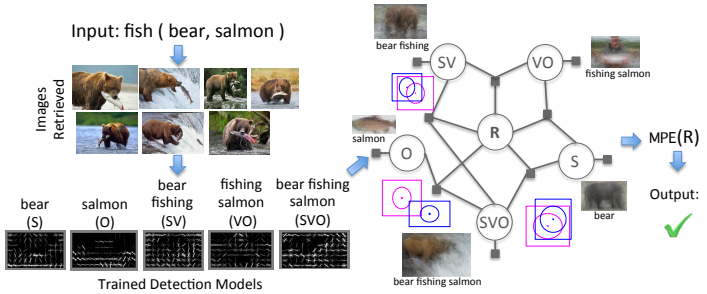

Figure 2: Approach Overview. Given a relation predicate, such as fish(bear,salmon) VisKE formulates visual verification as the problem of estimating the most probable explanation (MPE) by searching for visual consistencies among the patterns of subject, object and the action being involved.

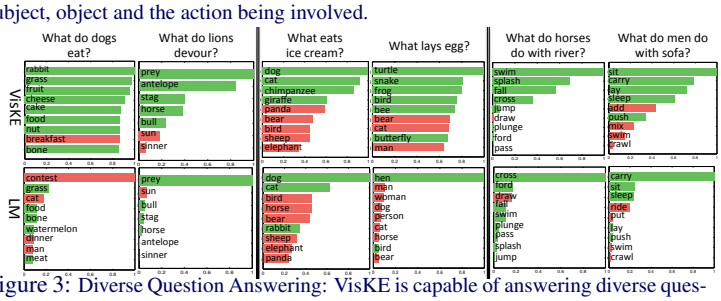

Figure 3: Diverse Question Answering: VisKE is capable of answering diverse questions about subjects, objects or actions. In comparison to the Language model [2], VisKE obtains richer and more precise answers to the questions.

	Base Set	Permute Set	Combined Set
Visual Phrase [3]	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 [2]	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 as well as the Combine set.

Relation Phrase Dataset and Evaluation: We gathered a new dataset of 'verb(subject, object)' relation phrases using Google Books Ngram [1]. We extracted 6093 relations from Google Books ('Base' set) and gathered new relations by randomly permuting the subjects, verbs and objects yielding and additional 6500 relations ('Permute' set). Table 1 summarizes the verification results on our relation phrase dataset using our approach compared with the Language Model [2] and other baselines.

Visual Question Answering: The visual knowledge that we have gathered using our visual verification method can help improve the reasoning within question-answering systems. In this paper, we show that our method enables answering generic questions like *what do dogs eat?*, *what animals lay egg?* (see Figure 3). Apart from answering generic questions, our approach can be useful for answering more specific questions such as those related to elementary-level general science questions.

[1] Alan Akbik and Thilo Michael. The weltmodell: A data-driven commonsense knowledge base. In *LREC*, 2014.

[2] J Huang et al. Exploring web scale language models for search query processing. In *WWW*, 2010.

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

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.