Visually-Verifiable Textual Entailment: A Challenge Task for Combining Language and Vision

Jayant Krishnamurthy
Allen Institute for Artificial Intelligence
2157 N. Northlake Way, Suite 110
Seattle, WA 98103
jayantk@allenai.org

Abstract

We propose visually-verifiable textual entailment as a challenge task for the emerging field of combining language and vision. This task is a variant of the well-studied NLP task of recognizing textual entailment (Dagan et al., 2006) where every entailment judgment can be made purely by reasoning with visual knowledge. We believe that this task will spur innovation in the language and vision field while simultaneously producing inference algorithms that can be used in NLP.

1 Introduction

It has long been acknowledged by the NLP community that extensive world knowledge and inference capabilities are necessary to perform basic language understanding tasks, such as reading a children's story (Minsky, 1975). Shallow knowledge representation techniques relying on only textual information have proven difficult to apply to complex inference problems because (1) much world knowledge is too obvious to be expressed in text, and (2) it is difficult to capture the complex structure of the real world within logical knowledge representations. Meanwhile, recent advances in computer vision have made it possible to train accurate object detectors (Russakovsky et al., 2014), suggesting that visual knowledge from images may be used to solve these natural language inference problems. However, many open problems must be addressed to successfully perform this combination, suggesting the need for a comprehensive challenge task to measure progress.

We propose that *visually-verifiable textual entailment* is a promising challenge task for combining language and vision. The task is to predict, given two texts, known as the text (T) and the hypothesis (H), whether the text *entails* the hypothesis ($T \vDash H$). T is said to entail H if, typically, a human reading T would infer that H is most likely true (Dagan et al., 2006). For example:

Text:	A man is flying a kite.
Hypothesis:	It is not raining

This example is an entailing pair because people typically do not fly kites in the rain. In visually-verifiable textual entailment, every entailment decision can be made purely on the basis of *visual* knowledge, i.e., knowledge that can be extracted from a large corpus of natural images. This criterion is satisfied by the above example – an image search for "man flying kite" returns no images where it is raining.

We believe that the task of visually-verifiable textual entailment is an exciting task for both the NLP and vision communities. From the NLP perspective, this task encourages the development of deep knowledge representation and inference techniques. These techniques may be able to solve more sophisticated inference problems than the shallow techniques – such as learning lexical substitution rules – currently in use (Giampiccolo et al., 2007). Recent work has also demonstrated the promise of using visual knowledge for entailment (Young et al., 2014). Furthermore, many NLP problems, such as coreference resolution and prepositional phrase attachment, can be posed as textual entailment problems; thus, this task provides a natural pathway for incorporating any developed techniques into downstream applications.

From the computer vision perspective, successfully performing this task requires developing accurate detection models of not just individual objects, but rather entire situations possibly unseen during training. The natural algorithm for visually-verifiable textual entailment is, given text T and hypothesis H, to first identify two sets of images, I_T and I_H, where the text and the hypothesis are true, respectively. Then, predict "en-

Proceedings of the 2015 Workshop on Vision and Language (VL'15), pages 1–3,
Lisbon, Portugal, 18 September 2015.

eating pizza		**eating spaghetti**		**eating an apple**	
holding pizza/a slice	3	enjoying spaghetti/meal	4	holding a fruit/apple	3
enjoying pizza	2	slurping spaghetti	2	thinking about things/apple	2
chewing pizza/food	2	holding a spoon/fork	2	posing with apple	2
consuming pizza	1	posing with spaghetti	1	biting apple	2

Table 1: Situation descriptions generated by Mechanical Turkers for three "eating" situations in preliminary data collection experiments. The descriptions are sorted by verb occurrence frequency.

tails" if $I_H \subseteq I_T$ and "not entails" otherwise.[1] Implementing this algorithm requires the ability to detect a wide variety of not just individual objects, but also attributes, relationships and events in images. Furthermore, it must be possible to compose these individual detectors in novel ways to form detectors for complete sentences. The variety problem has been partially addressed by webly-supervised algorithms for objects (Divvala et al., 2014; Chen et al., 2013) and subject-verb-object phrases (Sadeghi et al., 2015). The composition problem has also been examined, albeit with a very limited set of detectors (Matuszek et al., 2012; Krishnamurthy and Kollar, 2013). Progress on the proposed task requires improving on and combining these techniques.

2 Data Set

We propose to construct a data set for visually-verifiable textual entailment. As a starting point, we propose to focus on entailments between simple situations, given by a verb and optionally a subject and/or a direct object. This choice is motivated the fact that these situations are linguistically simple, yet can have complex entailments. For example, "eating an apple" $\models$ "holding an apple." However, "eating spaghetti" $\not\models$ "holding spaghetti;" rather "eating spaghetti" $\models$ "holding a fork." In the future, this data set can be expanded by including more complex language, e.g., prepositional modifiers.

To collect this data, we propose to use web image search and Mechanical Turk. First, we will manually identify a set of visual verbs and collect common arguments for them using a large corpus of syntactically parsed sentences. Combining these verb/argument pairs will produce a collection of situations. Second, we will feed these situations to an image search engine to retrieve multiple images depicting each situation. Third, we will construct a Mechanical Turk task for each image/situation pair, asking the worker to generate additional descriptions of the image. The design of this task will be tuned to generate more specific or general variants of the prompt situation (as in the example above). Because the generation occurs in the context of a particular image, not all of the generated situations will be entailed by the prompt situation. A final Mechanical Turk task will determine which situation pairs are entailments, thereby generating a data set with both positive and "near-miss" negative examples.[2]

We performed some preliminary experiments with this Mechanical Turk pipeline generating 18 situation descriptions for each of three "eating" phrases. The most frequent generations (sorted by verb) for each phrase are shown in Table 1. The resulting generations – though somewhat noisy – contain interesting structure: for example, both apples and pizza are held while being eaten. Apples are described with "biting," while spaghetti is described with "slurping."

3 Conclusion

We propose the task of visually-verifiable textual entailment as a challenge task for the field of combining language and vision. The object of this task is, given a text and a hypothesis, to predict whether the text entails the hypothesis. Crucially, the task design guarantees that each entailment decision can be made purely on the basis of visual knowledge. As a starting point, we propose to construct a data set of entailments between situations, i.e., verb/argument pairs, which appear to be the simplest case where nontrivial inference is required. Solving this entailment problem can require complex reasoning about real world situations, such as "eating pizza" $\models$ "holding pizza," whereas "eating spaghetti" $\not\models$ "holding a fork." We propose a data set collection methodology and present some preliminary data that demonstrates the potential of this task.

[1] This algorithm is unlikely to work in practice because it does not account for noise in the detections.

[2] If a binary yes/no entailment decision proves too ambiguous, we may also consider a ranking variant of the entailment task. In this variant, given a text and two hypotheses, the object is to predict which of the two hypotheses is more likely to be true.

Acknowledgements

We gratefully acknowledge Aria Haghighi, Oren Etzioni, Mark Yatskar and the anonymous reviewers for their helpful comments.

References

Xinlei Chen, Ashish Shrivastava, and Arpan Gupta. 2013. NEIL: Extracting visual knowledge from web data. In *2013 IEEE International Conference on Computer Vision (ICCV)*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising textual entailment*, pages 177–190. Springer.

Santosh K Divvala, Alireza Farhadi, and Carlos Guestrin. 2014. Learning everything about anything: Webly-supervised visual concept learning. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*.

Jayant Krishnamurthy and Thomas Kollar. 2013. Jointly learning to parse and perceive: Connecting natural language to the physical world. *Transactions of the Association for Computational Linguistics*, 1:193–206.

Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. 2012. A joint model of language and perception for grounded attribute learning. In *Proceedings of the 2012 International Conference on Machine Learning*.

Marvin Minsky. 1975. A framework for representing knowledge.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2014. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*.

Fereshteh Sadeghi, Santosh K Divvala, and Ali Farhadi. 2015. VisKE: Visual knowledge extraction and question answering by visual verification of relation phrases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.