Situation Recognition: Visual Semantic Role Labeling for Image Understanding

Mark Yatskar[1], Luke Zettlemoyer[1], Ali Farhadi[1,2]
[1]Computer Science & Engineering, University of Washington, Seattle, WA
[2]Allen Institute for Artificial Intelligence (AI2), Seattle, WA
[my89, lsz, ali]@cs.washington.edu

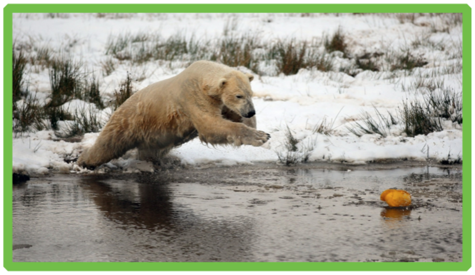
1. Supplementary Material

In this supplementary material, we discuss quality control methods we employed during data collection. Managing quality is difficult in our setting because of the complexity and scale of imSitu. Foremost, crowd workers vary in ability but also imSitu has over 1700 verb specific semantic roles. These roles may be ambiguously defined or an image might contain a situation where the semantic role values are challenging to determine . Our quality control algorithms models (1) the quality of a crowd worker, (2) the quality of a semantic role definition, and (3) the clarity of an image. These three elements are combined in a factorization method that assigns responsibility for low agreement between a pair of annotations. When the algorithm determines that crowd workers are the source of significant disagreement, they are warned and then banned.

1.1. Value Filling Quality Control

Quality control during the collection of imSitu addresses the challenge of assigning blame for high disagreement to either a poorly defined semantic role, a difficult image or a bad worker. For example, in Figure 1, the first and third workers produced identical annotations for all semantic roles except "part" while the second worker never agreed with other two. The target of our algorithm is to assign low quality to the second worker and the semantic role "part" and assign high quality to the image and all other semantic roles and workers.

For every worker, verb specific semantic role and image, we introduce real valued factor $f \in F$ that measures quality. We define a dataset D of pairs of annotations $(P, v, v') \in D$ where v and v' are values on the same semantic role of the same image but by different workers and $P \subset F$ is the set of factors involved in that pair. For example, in Figure 1, the first column would introduce three elements in D. For the element involving the first and second annotations, P would contain factors for the image and the semantic role "agent", the worker "turker1" and "turker2" and then v and v' would be "bear" and "deer" respectively. We assume a function Q that can take two values and measure their agreement on a scale of zero (low agreement) to one (high agreement). We minimize the following objective by assigning values to every $f \in F$:



JUMPING					
	AGENT	START	PART	END	PLACE
TURKER1	BEAR	SHORE	FOOT	STREAM	TUNDRA
TURKER2	DEER	-	WATER	-	-
TURKER3	BEAR	SHORE	AIR	STREAM	TUNDRA

Figure 1. Three hypothetical realized frames for an image of a bear jumping into a stream. Green and red indicate high and low quality, respectively. In this example, the second annotator is not accurately annotating the situation and the semantic role "part" is ambiguous so no annotator is able to accurately assign it a value.

$$\sum_{(P,v,v') \in D} ||Q(v, v') - \prod_{f \in P} f||_2 \qquad (1)$$

The minimization of this objective assigns a low value to at least one factor participating in pairs with low $Q(v, v')$ and high values for factors otherwise.

Equation 1 was minimized regularly during the data annotation process. The function Q was one over three plus the WordNet distance between v and v'. We found in practice this smoothly balanced the use of similar synsets. Factors were all initialized randomly around one, and we optimized with stochastic gradient ascent. A crowd worker that had a factor whose value was more than one standard deviation lower than other workers was warned and encouraged to email us for feedback. All workers who requested feedback had issues such as misinterpreting the meaning of semantic roles and overly general assignment of values. If the rating of a poor worker did not improve upon the submission of more annotations, they were automatically banned. Images and semantic roles with low values for factors are an opportunity to further improve the imSitu by examining reasons they were assigned lower quality but we leave this for future work.

1.2. Candidate Image Filtering Quality Control

Crowd workers filtering images retrieved from Google image search were also rated for quality, using a slight variation of the algorithm described above. In this filtering phase, crowd workers were required to select images that correspond to a verb from a set of 45 images. Factors were introduced for workers and the function Q measured the average number of images two workers agreed on in an identical set. Again, workers were warned if their quality scores were lower than one standard deviation than their peers and banned if they did not improve. Only images that were selected by at least two workers with a quality score greater than one standard deviation below the average were used for value filling.