Visual Abstraction for Zero-Shot Learning

Stanislaw Antol
Virginia Tech
santol@vt.edu

C. Lawrence Zitnick
Microsoft Research
larryz@microsoft.com

Devi Parikh
Virginia Tech
parikh@vt.edu

1. Introduction

Zero-shot learning (ZSL) [3, 1] involves training models for visual concepts without requiring any training images. Recent work utilizes textual descriptions (*e.g.*, attributes) for ZSL. This works well for categories that are easily *describable*, but it is unclear how to extend this work to ones that are not. For example, trying to describe the concept of two people dancing with each other (*e.g.*, bottom-right in Figure 2) via text or attributes would be cumbersome (at best), since you would need to describe the relative positions and angles of many body parts.

In this work, we introduce a novel modality for ZSL that utilizes visual abstraction. From a collection of clipart, users create illustrations depicting a concept. These illustrations are then used for training. We specifically focus on learning concepts involving individual poses and interactions between two people. We use an intuitive and simple interface (see Figure 1) that allows users to specify poses, expressions, and genders of people. Surprisingly, as demonstrated on two different datasets, our models learnt on visual abstractions are effective at classifying *real images*.

2. Datasets

To test our approach, we utilize two datasets:

INTERACT: We introduce this new dataset that focuses on *two people* interacting via different verb phrases. They include transitive verbs (*e.g.*, "A is pushing B"), joint activities (*e.g.*, "A is dancing *with* B"), movement verbs with directional prepositions (*e.g.*, "A is walking *to* B"), and posture verbs with locational prepositions (*e.g.*, "A is sitting *next to* B"). We combine different verbs with different prepositions to get 60 verb phrases, including ones that share a verb but contain different prepositions, such as "running *to*" and "running *away from*." The dataset was collected via the crowdsourcing service Amazon Mechanical Turk (AMT). We also used AMT to annotate the poses, eye gazes, expressions, and genders of people in the images. After some filtering (*e.g.*, duplicate removal), we have 3,172 images (examples in rightmost two columns of Figure 2).

PARSE: We also utilize the standard PARSE [4] dataset, which contains 305 images of *individuals* in various poses. We selected a *semantic subset* of 108 images by manually grouping them into 14 categories (*e.g.*, "is dunking", "is diving for an object"). We collected the same annotations as for INTERACT, except for pose (which was already available with the dataset).

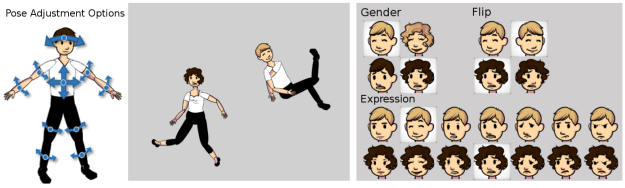

Figure 1. User interface (with random initialization) used to collect abstract illustrations on AMT. Workers are able to manipulate pose, expression, gaze direction, and gender.

3. Approach

We conjecture that our concepts of interest depend primarily on four factors: pose, eye gaze, facial expression, and gender. Our user interface is shown in Figure 1. Initially two people (one blonde-haired and one brown-haired) are shown with random poses, facial expressions, gaze directions (*i.e.*, "flip"), and genders. Since the variety of poses a person may assume is quite large, we allow our subjects to manipulate the poses (*i.e.*, joint angles and positions) of both people in a continuous space by dragging the various body parts. The facial expressions are chosen from seven prototypical expressions (the same selection was used for the real image annotations in Section 2). Finally, the subjects may horizontally flip the people to change their perceived eye gaze direction.

We collect training illustrations for ZSL via AMT. A user is prompted with a sentence describing one of the concepts to be learnt. Note that while we have semantic category names, *describing* the categories via text or attributes in order to build a visual model of them is not feasible. The user creates an illustration depicting the prompted concept using our interface (Figure 1). To promote diversity, we encouraged them to imagine any objects or background and ensure that the poses are consistent with the imagined scene (*e.g.*, they can imagine a chair and make someone sitting on it). A sample of these illustrations can be seen in the left five columns of Figure 2. When we collect illustrations for PARSE categories, the interface remains the same, except

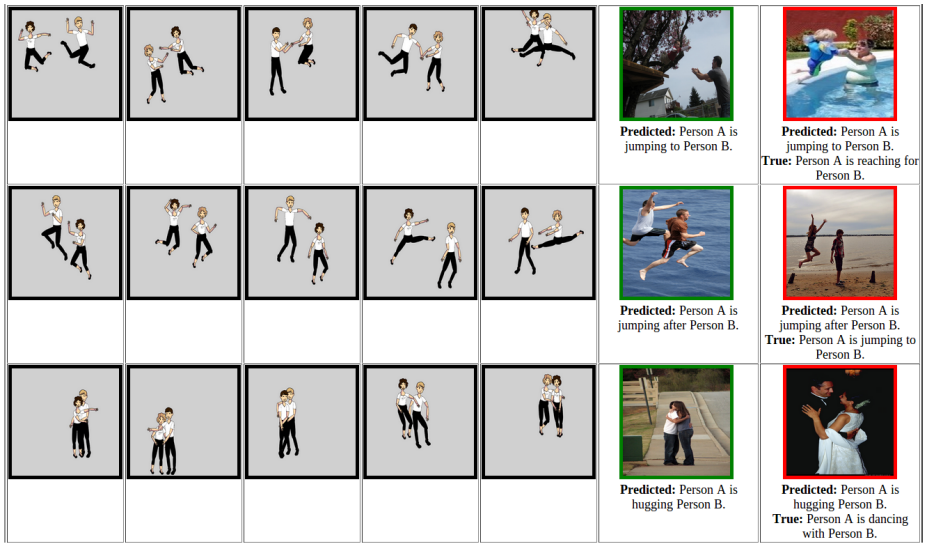

Figure 2. On the left, we show 5 random illustrations (of 50) used to train the classifiers, showing that workers typically do a reasonable job. The 6th and 7th columns contain the most confident true positive and false positive for a given category, respectively. Mistakes include choosing a semantically reasonable verb (top row), choosing the incorrect preposition (middle row), and incorrect prediction due to the pose similarity between two classes (bottom row).

that we remove one of the two clipart people.

In addition to gender, gaze, and expression features, we also extract contact features, joint angles, and general limb position features (*i.e.*, how far from the head, torso, and leg region the limbs are) from the 14 joints. This results in a total of 765 and 247 features for INTERACT and PARSE, respectively. The same set of features are extracted from real images using all of the annotations described in Section 2 or substituting ground truth poses with automatically detected ones [5]. We can now train classifiers on abstract illustrations and test them on real images.

4. Experimental Results

We use linear SVMs (liblinear [2]) with the cost parameter, C, set to 0.01 in our experiments. We evaluate average class accuracies over all 60 and 14 categories for INTERACT and PARSE, respectively as we increase the number of training illustrations per category. Results can be found in Figure 3. Each point on the curve is an average over 50 random selections of training illustrations. We see that even one illustration is able to perform several times better than random on both of our datasets. We can further improve performance by adding training illustrations, although we begin to saturate around 20 training examples. We also did a human agreement study on AMT for INTERACT. Across all images, out of 10 people, the correct verb phrase was only selected 52% of the time, which demonstrates how ambiguous this task is. Figure 2 shows some qualitative results.

We now briefly mention additional results whose details we can not show due to space constraints. We investigated how many mistakes are due to the classifiers incorrectly choosing the preposition or choosing the correct preposition but wrong verb. We also learn instance-level concepts (*e.g.*, corresponding to a specific real image in the database). Note that even naming such specific concepts via succinct text is not feasible. Hence to get training illustrations, we

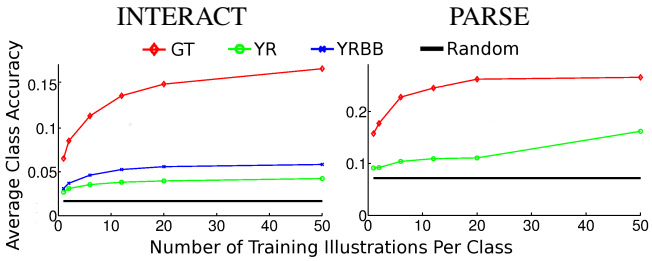

Figure 3. Initially, adding more training illustrations improves classification performance, but soon saturates. We show results using ground truth (GT), output from a pose detector (YR), and, for INTERACT, bounding box assisted pose detector (YRBB).

flashed a real image to AMT workers for 2 seconds and had them re-create the scene using our visual abstraction interface. This simulates a scenario where a user has a mental model for a specific concept, and depicts it using our visual abstraction to train the system for that mental concept. At test time, given such an illustration, nearest neighbor matching was used to identify the corresponding real image instance. We find that this performs significantly better than chance. Further, we learn a mapping between our abstract world and the real world using a Generalized Regression Neural Network. This further improves results.

References

[1] M. Elhoseiny, B. Saleh, and A. Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *ICCV*. IEEE, 2013. 1

[2] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *JMLR*, 2008. 2

[3] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*. IEEE, 2009. 1

[4] D. Ramanan. Learning to parse images of articulated bodies. In *NIPS*. 2007. 1

[5] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*. IEEE, 2011. 2