Learning semantic relationships for better action retrieval in images

Vignesh Ramanathan[1,2], Congcong Li[2], Jia Deng[3,2],
Wei Han[2], Zhen Li[2], Kunlong Gu[2], Yang Song[2], Samy Bengio[2], Chuck Rossenberg[2] and Li Fei-Fei[1]
[1]Stanford University, [2]Google, [3]University of Michigan

Human actions capture a wide variety of interactions between people and objects. As a result, the set of possible actions is extremely large and it is difficult to obtain sufficient training examples for all actions. However, we could compensate for this sparsity in supervision by leveraging the rich semantic relationship between different actions. A single action is often composed of other smaller actions and is exclusive of certain others. We need a method which can reason about such relationships and extrapolate unobserved actions from known actions. Hence, we propose a novel neural network framework which jointly extracts the relationship between actions and uses them for training better action retrieval models. Our model incorporates linguistic, visual and logical consistency based cues to effectively identify these relationships. We train and test our model on a largescale image dataset of human actions. We show a significant improvement in mean AP for action retrieval compared to different baseline methods including the HEX-graph approach from Deng et al. [1].

We build our action retrieval method on top of a standard retrieval model with ranking loss. For an action A, we wish to learn a weight vector w_A to minimize the following loss:

$$C_A = \sum_{\substack{I^+ \in \mathcal{I}_A \\ I^- \in \mathcal{I}_{\overline{A}}}} \max\left(0, 1 + w_A^T (f_{I^-} - f_{I^+})\right), \quad (1)$$

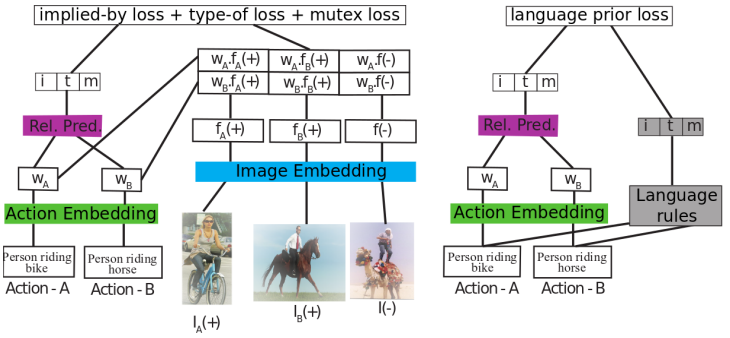
where $\overline{A}$ is the set of actions unrealted to A, and f_I is the feature representation of image I.

Next, given a pair of actions A and B, we wish to identify the relationship between them. The identified relationship is incorporated into our action retrieval model. These relationships determine the visual co-occurrence of actions within the same image. We define three kinds of relations following the recent work from [1]:

- **implied-by**: An action A is implied-by B, if the occurrence of action B implies the occurrence of A as well. This is similar to the *parent-child* relationship between A and B in a HEX-graph.
- **type-of**: An action A is a type-of B, if action A is a specific type of the action B. This is similar to *child-parent* relationship between A and B in a HEX-graph.
- **mutually exclusive**: An action A is mutually exclusive of B, if occurrence of A prohibits the occurrence of B.

We propose two novel objective functions which leverage visual information and logical consistency to identify these relationships between actions. The neural network component which uses visual information to determine relationships is shown in Fig. 1 (a). The relation between the actions is predicted by a tensor product layer on top of the weight vectors for the actions. The predicted relationship is represented by a 3 element vector $r_{AB} \in [0,1]^3$. As shown in the figure, we define a loss function corresponding to each of the three relations. We also use simple cues based on the text corresponding to the actions to establish a language-prior for these relationships (Fig. 1 (b)). For instance, we could identify that "Person riding animal" is implied-by "Person riding horse" due to the child-parent relation betwen "horse" and "animal" in WordNet.

Evaluation Most existing action datasets such as the PASCAL actions [2], as well as the Stanford-40 [4] are relatively small, with a maximum of 40 actions. The actions in the datasets were carefully chosen to be mutually exclusive of each other, making them less practical for real world settings. However, to demonstrate the efficacy of our method, we need a large dataset of human actions, where the actions are related to each other. Hence, we evaluate the performance of our model on a large dataset of 27425($27K$) actions obtained form Google image search. These actions are a subset of popular queries to the image search engine. This dataset was curated based on user clicks, to remove noisy examples for each action.



(a) Action retrieval with relationship (b) Language prior for relationship

Figure 1: The two main components of the relationship prediction model are shown. (a) defines a loss function which binds the predicted relationship with the learned action models and (b) regularizes the predicted relations with a language prior.

Method	27K mAP(%)	2.8K mAP(%)	Stan-81 mAP(%)
SOFTMAX	44.02	35.48	36.14
LANGRELWITHHEX [1]	-	37.12	36.48
RANKLOSS	46.43	35.56	36.38
DEVISE [3]	34.33	38.77	34.11
OURFULLMODEL	**54.78**	**45.82**	**38.73**

Table 1: Results of action retrieval on the 27K, 2.8K and an extended version of the Stanford-40 actions dataset.

We also run experiments under an additional setting of 2.8K actions, where we make the test images publicly available. In this setting, we use 2880 actions which form a subset of the 27K actions. However, we do not use a hand-curated training dataset with clean labels as before. Rather, while training the model, we treat the top 30 images returned by Google image search as ground truth positive images for each action, and the next 5 images are used for cross validation. Since the images are returned based on the text accompanying the images, the data could be noisy.

We also test our model and provide results for an extension of the Stanford 40 actions dataset. We relabel a set of images in the dataset into 41 additional labels. The results for these 3 datasets are shown in Tab. 1. Here, we compare our performance to the HEX-graph approach from Deng et al. [1] and the DeViSe model from [3].

[1] Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. Large-scale object classification using label relation graphs. In *Computer Vision–ECCV 2014*, pages 48–64. Springer, 2014.

[2] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[3] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, pages 2121–2129, 2013.

[4] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1331–1338. IEEE, 2011.

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