Learning semantic relationships for better action retrieval in images (Supplementary)

Vignesh Ramanathan*, Congcong Li†, Jia Deng+, Wei Han†,
Zhen Li†, Kunlong Gu†, Yang Song†, Samy Bengio†, Chuck Rossenberg† and Li Fei-Fei*

*Stanford University, †Google, +University of Michigan

vigneshr@cs.stanford.edu, congcongli@google.com, jiadeng@umich.edu

{weihan, zhenli, kunlonggu, yangsong, bengio, chuck}@google.com, feifeili@cs.stanford.edu

1. Experiment on Stanford 40 actions [5]

The original Stanford 40 actions dataset [5] has a carefully chosen set of 40 actions which are mutually exclusive of each other. This makes the dataset less applicable for large scale settings such as ours. Nevertheless, in order to demonstrate results on this dataset, we extend it with 41 additional action labels as explained below.

Setup We introduce 41 new action labels to this dataset. The additional actions are chosen such that they are *implied-by* one or more of the original 40 actions. The newly added labels are shown in Tab. 2. Against each of the original 40 actions, we show the set of newly added actions which are implied by this original action. We follow the experimental protocol form Deng et al. [2] and "relabel" a subset of the images to the newly added actions. More precisely, we relabel 50% of the images belonging to an original action to one of the newly introduced actions which is *implied-by* this original action (as shown in Tab. 2). For instance, some images belonging to "playing violin" are now relabelled to "playing an instrument". We do this for both the training and testing images. Note that we do not add any new images to the dataset, and each image still has exactly only one label. Hence, the original set of 4000 training images are now redistributed into 81 classes.

Evaluation We use mean average precision to evaluate our method as before. Since the newly added actions are related to each other, the positive image of an action could also be a positive for other actions. Hence, for every action we only treat the images of other actions which are mutually exclusive or unrelated as negative examples.

Experiment We use the same deep neural network architecture as before. We initialize the relation prediction tensor layer as well as the image embedding layer with the corresponding layers learned from the $27K$ action dataset. We use the same hyper parameters as before.

Results We show results on our extended version of the Stanford 40 actions dataset in Tab. 1. Additionally, we also separately list the results for the newly added action labels. The baselines are the same as explained in the main draft.

Method	mAP (%) 81 ac.	mAP (%) 41 new ac.
LOGISTIC	21.85	18.23
SOFTMAX	36.14	33.19
LANGRELWITHHEX [2]	36.48	32.77
RANKLOSS	36.38	31.72
DEVISE [3]	34.11	30.13
OURONLYLANGREL	37.12	34.23
OURFULLMODEL	**38.91**	**37.22**

Table 1. Results of action retrieval on the extended version of the Stanford 40 actions dataset. The first column shows results for all the 81 actions, while the second column shows results for only the 41 newly added actions.

Our full model outperforms all baseline models on the 81 actions. The performance improvement is more pronounced for the newly added action labels shown in the second column. The added actions are implied-by the original actions, and identifying these implied-by relationship would lead to better performance gain as explained in the main draft. As expected, the improvement in mean AP for these newly added actions is seen to be larger than that for the original 40 actions.

2. Language based rules for action relations

We discuss the simple set of rules which are used to determine the relationship between a pair of actions whenever possible. These rules are based on WordNet relationship between entities. The actions in our dataset are of one of the following forms:

- *SVD* ⟨subject, verb, direct-object ⟩. eg: Person eating food
- *SVP* ⟨subject, verb, prepositional-object ⟩. eg: Person eating with fork

978-1-4673-6964-0/15/$31.00 ©2015 IEEE

Original action	Newly added *implied-by* actions
applauding	clapping , cheering others
blowing bubbles	blowing something , holding a container , releasing something
brushing teeth	holding a toothbrush , doing an oral activity
cleaning the floor	cleaning something , holding a cleaning device , working on the floor
climbing	clinging to something , doing hard physical activity , playing some sport
cooking	preparing food , handling food
cutting trees	cutting something , holding a cutting tool
cutting vegetables	preparing food , handling food , cutting something , holding a cutting tool
drinking	holding a container
feeding a horse	handling food , interacting with an animal
fishing	doing something near water , holding something
fixing a bike	fixing something , working with a bike , working with a vehicle
fixing a car	fixing something , working with a vehicle , working with a car
gardening	working on the ground , tending to a garden , doing an outdoor activity
holding an umbrella	holding something
jumping	–
looking through a microscope	looking through something , bending over something
looking through a telescope	looking through something
playing guitar	playing an instrument
playing violin	playing an instrument
pouring liquid	holding a container
pushing a cart	doing hard physical activity , pushing something
reading	looking at something , looking at a book
phoning	interacting with the phone , holding something
riding a bike	working with a bike , working with a vehicle
riding a horse	interacting with a horse , interacting with an animal
rowing a boat	doing hard physical activity , doing something near water
running	doing hard physical activity , playing some sport
shooting an arrow	playing some sport , releasing something
smoking	holding something , blowing something , doing an oral activity
taking photos	holding something , looking into something
texting message	interacting with the phone , holding something , typing on something
throwing frisby	releasing something , playing some sport
using a computer	looking at a screen , typing on something
walking the dog	moving
washing dishes	cleaning something
watching TV	looking at a screen
waving hands	–
writing on a board	writing on something , holding something
writing on a book	writing on something , holding something

Table 2. The original 40 actions of the Stanford dataset [5] are shown in the first column. The newly added action labels are shown in the second column. Again each original action, we show the subset of newly added actions which are *implied by* this original action. For the experiments, 50% of the images belonging to an original action is relabelled to one of its *implied by* actions shown in the second column.

- *SVDP*⟨subject, verb, direct-object, prepositional-object ⟩. eg: Person eating food with fork.

Given these forms, we use the following rules for determining relationship between actions A_1 and A_2, where the earlier rules take precedence over the later rules in case of conflict. These rules are a direct consequence of the relationships defined in WordNet, and is similar to the hierarchy based structure used in other works such as [4].

1. A_1 **is *implied-by*** A_2, if A_1 has *SVD* form, A_2 has *SVD* or *SVDP* form and all the three words of A_1 are either synonyms, meronyms, hyponyms of the corresponding words in A_2. eg: "Person cleaning building" is *implied-by* "Woman washing window"

2. A_1 **is *implied-by*** A_2, if A_1 has *SVP* form, A_2 has *SVP* or *SVDP* form and all the three words of A_1 are either synonyms, meronyms, hyponyms of the corresponding

words in A_2. eg: "Person drinking from container" is *implied-by* "Person drinking water from bowl"

3. A_1 **is *type-of*** A_2, if A_1 has *SVD* or *SVDP* form, A_2 has *SVD* form and the subject, verb, direct-object of A_1 are either synonyms, holonyms, hypernyms of the corresponding words in A_2. eg: "Chef baking pizza in oven" is *type-of* "Person cooking food"

4. A_1 **is *type-of*** A_2, if A_1 has *SVP* or *SVDP* form, A_2 has *SVP* form and the subject, verb, prepositional-object of A_1 are either synonyms, holonyms, hypernyms of the corresponding words in A_2. eg: "Teacher writing on board with chalk" is *type-of* "Person writing with something"

5. A_1 **is *mutually exclusive* of** A_2, if A_1 has *SVD* form, A_2 has *SVD* or *SVDP* form, exactly two words of A_1 are either synonyms, hyponyms, hypernyms of the corresponding words in A_2, and the third word of A_1 shares a common hypernym with the corresponding word of A_2. eg.: "Perosn riding horse" is *mutually exclusive* of "Woman riding camel with a hat"

6. A_1 **is *mutually exclusive* of** A_2, if A_1 has *SVP* form, A_2 has *SVP* or *SVDP* form, exactly two words of A_1 are either synonyms, hyponyms, hypernyms of the corresponding words in A_2, and the third word of A_1 shares a common hypernym with the corresponding word of A_2. eg.: "Woman eating on table" is *mutually exclusive* of "Person eating food on floor"

7. A_1 **is *mutually exclusive* of** A_2, if A_2 is mutually exclusive of A_1.

3. Disallowed states for consistency loss

As explained in the main draft, certain sets of relationships between a triplet of actions are deemed to be inconsistent with each other. We penalized these relationships in the consistency loss. We list these inconsistent relationships in Tab. 3. For instance, the first row provides the following inconsistent relationship: action A_1 is *implied-by* A_2, A_1 is *type-of* A_3 and A_2 is *mutually exclusive* of A_3.

4. Implementation details

The full objective is minimized through downpour stochastic gradient descent [1]. The various hyper-parameters of the model: $\{\beta, \lambda, \alpha_r, \alpha_c, \alpha_n\}$, were obtained though grid search to maximize performance on a validation set. These parameters were set to $1000, 0.01, 5, 0.1, 10$ respectively for both experimental settings in the next section. The embedding dimension n was set to 64. While training the model, we run the first few iterations without the relation prediction objectives. We provide more details in the supplementary material.

A_1-A_2	A_1-A_3	A_2-A_3
i	t	m
i	t	i
t	i	m
t	i	i
t	t	m
i	m	i
i	m	t
m	i	i
m	i	t
t	m	t
m	t	i

Table 3. The set of inconsistent relationships are shown for a triple of actions A_1, A_2, A_3. The first column denotes the relationship of A_1 with respect to A_2, the second column denotes the relationship of A_1 with respect to A_3 and the third column denotes the relationship of A_2 with respect to A_3. Here, "p" denotes *implied-by*, "t" denotes *type-of* and "m" denotes *mutually exclusive* relationships.

We also observed a performance gain by fixing the relation predictions and only optimizing the action prediction objective in the final few iterations.

We use a batch size of 8 actions for the action recognition model, where each action is accompanied by 1 positive and 7 negative images, leading to a total of 128 images per batch. Similarly, we use a batch size of 10 action pairs for the relationship prediction models, where each action pair is accompanied by 12 images, corresponding to 4 positive images of each action and 4 negative images. We initialize the learning rate at 0.1 and gradually decrease it during training based on a visual inspection of the cost curve.

References

[1] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, A. Senior, P. Tucker, K. Yang, Q. V. Le, et al. Large scale distributed deep networks. In *Advances in Neural Information Processing Systems*, pages 1223–1231, 2012.

[2] J. Deng, N. Ding, Y. Jia, A. Frome, K. Murphy, S. Bengio, Y. Li, H. Neven, and H. Adam. Large-scale object classification using label relation graphs. In *Computer Vision–ECCV 2014*, pages 48–64. Springer, 2014.

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

[4] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2712–2719. IEEE, 2013.

[5] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. 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.