# Learning semantic relationships for better action retrieval in images

Anonymous CVPR submission

Paper ID 0195

## Abstract

*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 theses relationships. We train and test our model on a new largescale image dataset of human actions under two settings with $27K$ and $2K$ actions. We show a significant improvement in mean AP compared to different baseline methods including the state-of-the-art HEX-graph approach from Deng et al. [8].*

## 1. Introduction

Humans appear in majority of visual scenes, and understanding their actions is the basis of successful human computer interaction. While action retrieval poses the same challenges as object recognition, one key difference is that the semantic space of actions is much larger. As shown in Fig. 1, actions are compositions of objects and there are many possible interactions even between the same set of objects. The distribution of objects in images is already long tailed; consequently actions would be distributed in a much more skewed way since most object combinations are quite rare. Thus for successful action retrieval, one has to address the fundamental challenge of learning with few examples. In the current work, we learn action models for retrieving images corresponding to a large number of human actions in this challenging setting.

An action such as "person interacting with panda" yields very few relevant results on image search. Can we still learn a reliable model with such sparse supervision? As shown in Fig. 1, the answer lies in the key observation that action classes are related to each other. We may have few instances for this action, but we have also seen "person feeding a panda", "person holding animals" etc. and we understand how these actions are semantically related. Thus we can readily extrapolate to recognize "person interacting with panda".

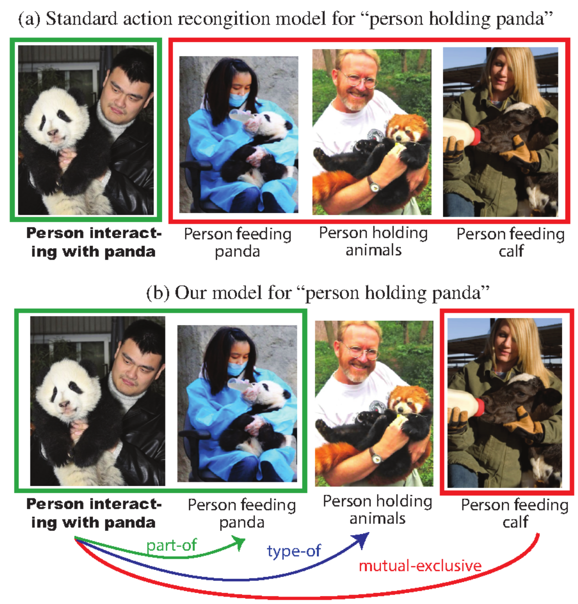


Figure 1. Given a query, such as "Person interacting with panda" (a) standard models for action recognition treat every action independently, while (b) our method identifies the relation between actions, and uses these relations to extrapolate labels for images of related actions. In this example, "person interacting with panda" is part-of "person feeding panda", and mutually exclusive of "Person feeding a calf". Hence, the images of these actions could also be used to train a model for "person interacting with panda". The green and the red boxes indicate the positive and negative examples considered by the methods for training the model.

This observation naturally leads to the idea of using a semantic graph that encodes relationship between classes. In fact, this idea was explored in the HEX-graph approach of Deng et al. [8]. However, their method left a key issue unaddressed: where does the graph come from in the first place? The experiments of [8] only used single entity classes and adapted WordNet[26] to heuristically obtain a HEX-graph for the entities. However, there is no such preexisting hierarchical structure for composite classes like *actions*.

To address this problem, we would like to automatically learn the semantic relations between actions. This cannot be simply circumvented by crowdsourcing. It would be prohibitively expensive to manually annotate relations even between every pair of object-verb-object triplets, leave alone actions. On a more fundamental level, we would also like computers to be able to automatically extract knowledge from data. *The main contribution of our work is a new deep learning framework which unifies the two problems of learning action retrieval models and predicting action relationships.* To the best of our knowledge, this is the first such attempt for retrieval of human actions.

We leverage two key insights to build our model, along with the known fact that semantic relations help training visual models:

1. Some relations can be deduced from linguistic sources. Automatic relationship prediction in NLP [4, 24] is far from perfect. Nevertheless, linguistic tools such as WordNet still provide valuable cues. As an example, the parent-child relationship between "panda" and "animal" tells us that "Person holding panda" is part-of "Person holding animals".

2. Relationship between actions like "feeding a panda" and "interacting with a panda" Fig. 1 cannot be captured solely through language. The visual knowledge from the action retrieval models could help us in such examples. Additionally, the logical consistency between actions can also be used to extrapolate new relations from existing ones. If we know "person feeding calf" excludes the action of "person feeding panda", and "feeding" is a type of "interaction", then we can infer that "person interacting with panda" is also exclusive of "person feeding calf".

We train our model on a large-scale dataset of 27425 actions collected by crawling the web for images corresponding to these actions. We show significant improvement compared to a standard recognition model, as well as the HEX-graph based approach from [8]. Additionally, we also provide results for a subset of 2000 actions, whose data is made publicly available.

## 2. Related work

**Semantic hierarchy for vision** In the last few years, different works [7, 25, 47, 11, 44, 16, 37, 9, 27, 1, 8] have tried to use preexisting structure between labels to train better models for image classification, and object segmentation [21]. Most related to our work is the recent work from Deng et al. [8], who use DAG relationships and mutual exclusions among entity labels to train better classifiers. All these works achieve a gain in performance, when provided with a fixed semantic hierarchy between labels. Such straightforward semantic relationships are absent for real world human actions. Hence, we automatically discover these relations.

Another line of work shares data between visually similar classes by learning grouping of class labels [31, 23, 22, 38, 3, 30, 17, 29, 45]. These methods typically cluster the labels or organize them in a hierarchical taxonomy based on visual similarity and co-occurrence. However, we learn semantic relationships based on both language and visual information, and we do not restrict ourselves to a hierarchical taxonomy.

**Building visual knowledge** Recently, there has also been a push in works such as [2, 46] to learn visual relationship between entity labels by mining images from the web. In particular, NEIL [2] extracts relationship between objects, attributes and scenes only based on the visual overlap between the corresponding images. They use the extracted relations as additional context for re-scoring objects and scenes. In contrast, we learn relationship between actions by minimizing a joint objective across all actions, and simultaneously learn models for action retrieval. Further, we provide a single neural network architecture to achieve this.

**Action recognition** Action recognition in images has been widely studied in different works such as [42, 15, 28, 41, 32]. They focus on improving performance for a small hand-crafted dataset of mutually exclusive actions such as the PASCAL actions and Stanford 40 actions [10, 43]. Most methods [42, 15, 28] try to improve the detection of objects or poses specific to these datasets, and are not scalable to larger number of actions. More recently, video action recognition [39, 33, 19] models have been quite successful for larger datasets such as UCF-101 [36], and the Sports-1M [19]. At this scale, the datasets are still composed of mutually independent actions such as sports activities. However, we focus on an almost open world setting for actions with rich semantic relationship between the actions.

**Joint image and text embeddings** Another class of work [12, 35, 18] tries to learn models in an open world setting by embedding textual labels, and images in a joint space. They learn a single embedding space, where text and associated images are close to each other. These methods only rely on textual similarity between sentences/words to capture visual similarity. Most of these methods treat sentences without textual overlap such as "drinking coffee" and "holding cup" to be dissimilar. Also, these methods are not constructed to handle asymmetric relations between classes. On the other hand, we explicitly learn asymmetric visual relationship between actions in our dataset.

**Relationship prediction in NLP** Our work also draws inspiration from research in NLP such as entailment[24] and

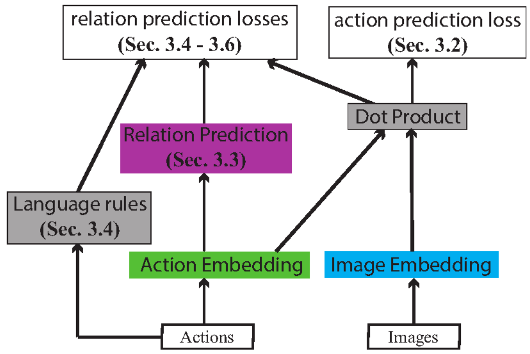


Figure 2. A schematic overview of our model for jointly predicting the relationship between actions, and learning action retrieval models.

natural logic [4]. In particular, our work is related to [34] which proposes a neural tensor layer to learn relationship between embeddings of textual entities.

# 3. Our approach

We wish to learn action retrieval models for a large number of actions which are related to each other. To learn good models, we would ideally like to have all action labels for all images in our dataset. In practice, obtaining multiple labels for an image does not scale with the number of actions and we are restricted to one label per image. However, if we understand the semantic relationship between different human actions, we can easily extrapolate missing labels from a single action. For example, we expect an image depicting "Person riding horse", to contain other actions such as "Person sitting on animal", "Person holding a leash" and to not contain "Person riding a camel".

Identifying such relationships is a challenging task in itself. While language can help to certain extent, we also need to use visual information to reliably identify relationships. The problems of training action retrieval models, and predicting relationships are closely coupled with each other. The main contribution of our work is a neural network architecture which can jointly handle these tasks.

A schematic of our model is shown in Fig. 2. Actions and images are embedded into vectors by embedding layers, and the relationship between actions are predicted from the action embeddings. We finally have a joint objective for learning action models and ensuring good relationship prediction. The objective has two main components[1]:

- Action prediction loss visualized in Fig. 3.
- Relation prediction loss composed of three modules, where each module is designed to capture a specific aspect of the relationship as shown in Fig. 4.

[1] While the loss functions are minimized jointly, we have shown them separately in the figures for the convenience of easy visualization.

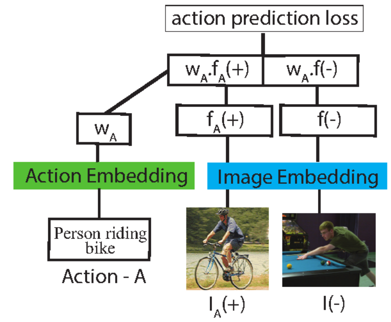


Figure 3. The action retrieval model, where the image and action embedding layers are shared with the modules in Fig. 4

## 3.1. Problem setup

We are given a set of actions $\mathcal{A}$, and for every action $A$ in $\mathcal{A}$ we have a set of positive images $\mathcal{I}_A$. We are also provided a set of related actions $\mathcal{R}_A \subset \mathcal{A}$, for every action $A$. For each action we wish to learn models which ranks the positive images of the action higher than the negative images. We also identify the relationship between $A$ and every action in $\mathcal{R}_A$. We obtain $R_A$ by selecting the actions whose top 100 images returned by Google image search have an overlap with those of the action $A$.

All the actions in our dataset contain one or both of the two structures: 1. $\langle$ subject, verb, object $\rangle$, eg.: "Person riding a horse" 2. $\langle$ subject, verb, prepositional object $\rangle$, eg.: "Person walking with a horse" This is a reasonable representation for actions as noted in past works such as [13].

## 3.2. Action retrieval

We first develop a basic action retrieval model (Fig. 3) which is later integrated with relationship prediction modules in the next few sections. We use a simple feed-forward architecture, where each action description $A$ from the set of actions $\mathcal{A}$ is represented by a weight vector $w_A \in \mathbb{R}^n$, and each image $I$ is represented as a feature vector $f_I \in \mathbb{R}^n$, and $n$ is the embedding dimension. The feature $f_I$ is obtained through a linear projection of the Convolutional Neural Network (CNN) feature, obtained from the last fully connected layer of a CNN architecture [20, 40]:

$$f_I = W_{im}\text{CNN}(I) + b_{im}, \tag{1}$$

where $\text{CNN}(I)$ represents the CNN feature of image $I$. The projection parameters $W_{im}, b_{im}$ are learned in the model. We assume that the actions which are not part of the set $\mathcal{R}_A$ are unrelated to $A$, and the corresponding images are treated as negatives. The action weight vector should assign a higher score to a positive image as compared to negatives. Hence, we define a ranking loss:

$$C_{ac} = \sum_{A} \sum_{\substack{I^+ \in \mathcal{I}_A \\ I^- \in \mathcal{I}_{\overline{A}}}} \max\left(0, 1 + w_A^T (f_{I^-} - f_{I^+})\right), \tag{2}$$

where $\overline{A} = \mathcal{A} \setminus \mathcal{R}_A$ is the set of actions unrelated to $A$.

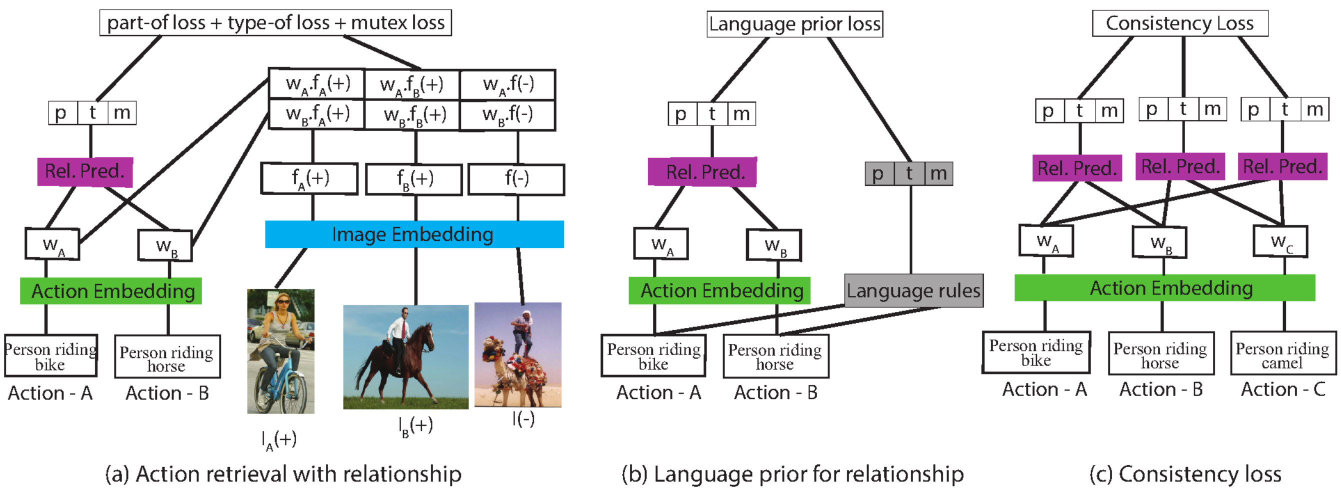


Figure 4. The different components of the relationship prediction model are shown, where the image and action embedding layers are shared with Fig. 3. (a) defines a loss function which binds the predicted relationship with the learned action models, (b) regularizes the predicted relations with a language prior, and (c) tries to enforce logical consistency between predicted relations.

### 3.3. Relationship prediction

Given a pair of actions $A$ and $B \in \mathcal{R}_A$, we wish to identify the relationship between them. These relationships determine the visual co-occurrence of actions within the same image. Naturally, we want to predict relations based on some visual representation of the actions. Hence, we formulate a relation prediction function on top of the action embeddings defined in the previous section. However, we first need a reasonable definition for relationship. We follow the recent work from [8] to define three kinds of relations:

- **part-of**: An action $A$ is part-of $B$, if the occurrence of action $A$ implies the occurrence of $B$ as well. This is similar to the *parent-child* relationship between $A$ and $B$ in a HEX-graph.
- **type-of**: An action $A$ is 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 denote the relationship by a binary vector $r_{AB} = [r^p_{AB}, r^t_{AB}, r^m_{AB}] \in [0,1]^3$, where $r^p, r^t, r^m$ denote *part-of*, *type-of* and *mutually exclusive* relationship values respectively. The relationship is predicted through a neural tensor network layer similar to the knowledge base completion work from Socher et al. [34]. This layer is followed by softmax normalization, as shown in Fig. 4. The predicted relationship can be written as:

$$r_{AB} = \text{softmax}_\beta\left(w_A \otimes W_{rel}^{[1:3]} \otimes w_B + b_{rel}\right), \qquad (3)$$

where the tensor $W_{rel}^{[1:3]} \in \mathbb{R}^{n\times n\times 3}$ and $b_{rel} \in \mathbb{R}^3$ are the parameters to be optimized, and $\text{softmax}_\beta : \mathbb{R}^3 \mapsto \mathbb{R}^3$ is the softmax normalization function with parameter $\beta$.

### 3.4. Language prior for relationship

As noted in the introduction, the text of an action carries valuable information about its relations. However, predicting relations between any two generic textual phrases is a rather challenging problem in NLP [4, 24]. The performance of such systems is often unsatisfying for use in higher level tasks such as ours. We propose to get around this limitation by capitalizing on the structured nature of actions in our problem. We define a set of simple rules based on WordNet hierarchies to impose a prior on the relationship between some of the actions in our dataset. If none of the rules are satisfied, we do not use any prior, and let the other components of the model decide the relationship. Some rules used in our system are visualized in Fig. 5. The complete set of rules are provided in the supplementary.

It is important to note that these rules are not always accurate, and can be quite noisy as shown in the third example of Fig. 5. Further, the rules are not satisfied for a large number of cases. We observed that $41.69\%$ of the relationships in our datasets do not satisfy the listed language based rules. Hence, the relationship set by these rules should only be treated as a noisy prior, and cannot be directly used to combine data as we show later in the experiments as well.

We use the relationship prior from these rules to define a loss function as shown in Fig. 4(b). If the NLP prior for the relationship is given by the vector $\tilde{r}_{AB}$, then we define an $\ell_1$ loss function as follows:

$$C_{nlp} = \sum_{A} \sum_{B\in\mathcal{R}_A} |r_{AB} - \tilde{r}_{AB}| \qquad (4)$$

### 3.5. Action retrieval with relationship

So far, we have defined a relation prediction layer and determined a language based prior for a subset of the rela-

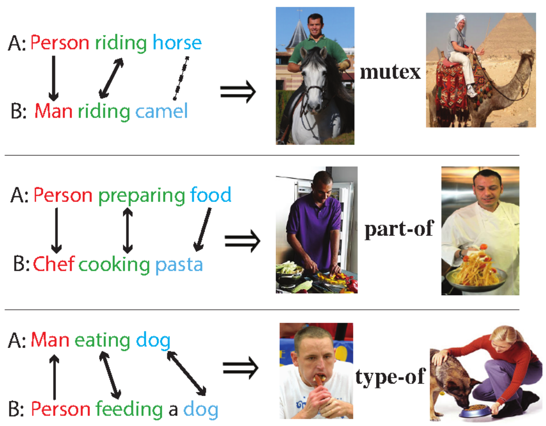


Figure 5. Some sample rules in our language prior are visualized here. These rules are derived from WordNet; the arrows represent parent-child relation in WordNet, and the dashed line corresponds to siblings. For instance, the first example implies that if the subjects are related as parent-child, the verbs are synonyms and the objects are siblings, then the actions are mutually exclusive. As seen in the third example, some relations derived can still be noisy due to lack of contextual information for the action.

tions. However, to fully use relationships for training better models, we still need to extrapolate relations which do not have a language prior. We propose two novel objective functions which leverage visual information and logical consistency to determine good action relationships.

**Visual objective** As mentioned earlier in the introduction, the relationship between actions determine how their training data can be shared between them. In particular, we define a specific loss function for each relation:

- If action $A$ is part-of $B$, the weight vector $w_A$ should rank the positive images of $B$ higher than the negatives of $A$, which in turn implies a small value for:

$$C^p_{AB} = \sum_{\substack{I^b \in \mathcal{I}_B \\ I^- \in \mathcal{I}_{\overline{A}}}} \max\left(0, 1 + w_A^T(f_{I^-} - f_{I^b})\right) \quad (5)$$

- If $A$ is type-of $B$, the weight vector of $w_B$ should rank the positive images of $A$ higher than negatives of $B$. Hence, we expect a small value for the cost:

$$C^t_{AB} = \sum_{\substack{I^a \in \mathcal{I}_A \\ I^- \in \mathcal{I}_{\overline{B}}}} \max\left(0, 1 + w_B^T(f_{I^-} - f_{I^a})\right) \quad (6)$$

- If $A$ is mutually exclusive of $B$, the weight vector $w_A$ should rank positive images of $A$ higher than the positives of $B$. Hence, we expect a small value for:

$$C^m_{AB} = \sum_{\substack{I^a \in \mathcal{I}_A \\ I^- \in \mathcal{I}_B}} \max\left(0, 1 + w_A^T(f_{I^b} - f_{I^a})\right) \quad (7)$$

Now, we combine these losses along with the corresponding relation prediction values to formulate an objective $C_{rec}$ as follows. The module of the neural network corresponding to this objective is shown in Fig. 4(a).

$$C_{rec} = \sum_{\substack{A \in \mathcal{A} \\ B \in \mathcal{R}_A}} r^p_{AB} \cdot C^p_{AB} + r^t_{AB} \cdot C^t_{AB} + r^m_{AB} \cdot C^m_{AB} \quad (8)$$

If the action weight vectors $w_A, w_B$ are properly trained, the loss function corresponding to the best relation would be small, causing the model to automatically choose the right relation. Similarly, if the relationship is chosen correctly, the training data of the actions would be correctly augmented, leading to better action weights.

**Consistency objective** We use logical consistency among the predicted relations as an additional cue to constrain the relationship assignment between actions. However, a global consistency constraint would span all action pairs and couple their relation predictions. To get around this problem, we propose a consistency cost only over triplets of related actions. We observe triplets of actions, and down weight inconsistent binary relationships between all pairs of actions in this triplet. For instance, we want to avoid inconsistent relationships such as: $A$ *is part-of* $B$, $B$ *is part-of* $C$ and $A$ *is mutually exclusive of* $C$. It is straight-forward to list out all the disallowed relationships for a triplet of actions (shown in the supplementary material). We refer to this set of disallowed relationships as $\mathcal{D} \subset \{p, t, m\}^3$, and define the consistency objective as follows:

$$C_{cons} = \sum_{\substack{A \\ B \in \mathcal{R}_A \\ C \in \mathcal{R}_B}} \sum_{d \in \mathcal{D}} r^{d_1}_{AB} \cdot r^{d_2}_{BC} \cdot r^{d_3}_{CA}, \quad (9)$$

where the disallowed relationship triplet $d$ is of the form $(d_1, d_2, d_3)$. The component of the neural network implementing this loss function is shown in Fig. 4(c).

### 3.6. Full model

We tie together the action prediction loss and the relation prediction losses in one single objective as shown below:

$$C = C_{ac} + \alpha_r C_{rec} + \alpha_n C_{nlp} + \alpha_c C_{cons} + \lambda \|W\|_2^2, \quad (10)$$

where $\alpha_r, \alpha_n, \alpha_c$ are hyper-parameters. The weights in the model $W = \{W_{im}, \bigcup_{A \in \mathcal{A}} w_A, W_{rel}\}$ are $\ell_2$ regularized with a regularization coefficient $\lambda$.

**Implementation details** The full objective is minimized through downpour stochastic gradient descent [5]. 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.


Figure 6. A few actions from our dataset along with images. For every action, we also show a sample related action. The relation from language prior is shown in red, and the correct relation predicted by our full method is shown in green.

# 4. Experiments

We evaluate the action retrieval performance of our model against different baselines under two experimental settings. We also present a detailed analysis of the relations learned by our model.

## 4.1. Dataset

As listed in Guo et al. [14], most existing action datasets such as the PASCAL actions [10], as well as the Stanford-40 [43] 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. They have very little or no overlap between the actions. 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 construct a dataset of 27425 action descriptions with very few restrictions on the choice of actions.

We present results on two different settings corresponding to $27K$ and $2K$ actions as explained below, where the data is made publicly available for $2K$ actions.

**27K actions**: We collected a set of positive examples for each action description by scraping the top results returned by Google image search. This dataset was curated by annotator ratings, to remove noisy examples for each action. Two thirds of the images per action are used for training, while the remaining images are held out for use in testing and validation. We treat 13700 actions and the associated held-out images as the validation set. The held-out images of the remaining 13725 actions are used for testing. We have $15 - 200$ training images per action resulting in a total of 910775 training images.

**2K actions**: We also run experiments under an additional setting, where we make the test images publicly available. In this setting, we use 2000 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. Nevertheless, as observed in Dean et al. [6], they contain sufficient information to train visual classifiers. Some sample actions and relations in our dataset are shown in Fig. 6. It is to be noted that the *test set* corresponding to the $2K$ actions is still curated with annotator ratings to remove noisy examples, and has no overlap with the training and validation data.

**Evaluation criteria** We use mean Average Precision (mAP) to evaluate our method in an image search setting, where we wish to retrieve the correct images corresponding to an action label from the test set. Note that, each test image could be associated with more than one correct action label due to the relationship between different actions in our dataset. However, we do not have the label corresponding to all actions for all images in the test set. Hence, for the sake of correct evaluation we also annotate a set of negative images for each action description and compare the scores of the true positives of an action with these annotated negatives for the action. Our test set typically contains 500 negative images and $3 - 10$ positive images for each action-label.

## 4.2. Results

We compare with the joint image-text embedding method from DeVise [12], as well as the recent HEX-graph method of using relations, proposed in [8]. The different baselines used for comparison are listed below:

1. LOGISTIC Model without relations, trained with logistic loss
2. SOFTMAX Model without relations, trained with softmax loss
3. LANGRELWITHHEX Action recognition model trained with the HEX-graph loss function proposed in [8]. Only the relations from Language prior are used to to construct a HEX-graph.
4. OURRELWITHHEX We use the relationships learned by our full model in the HEX-graph loss function. Since the HEX-graph needs to be consistent without cycles we first build a Maximum Spanning Forest (MSF) based on our learned relations.

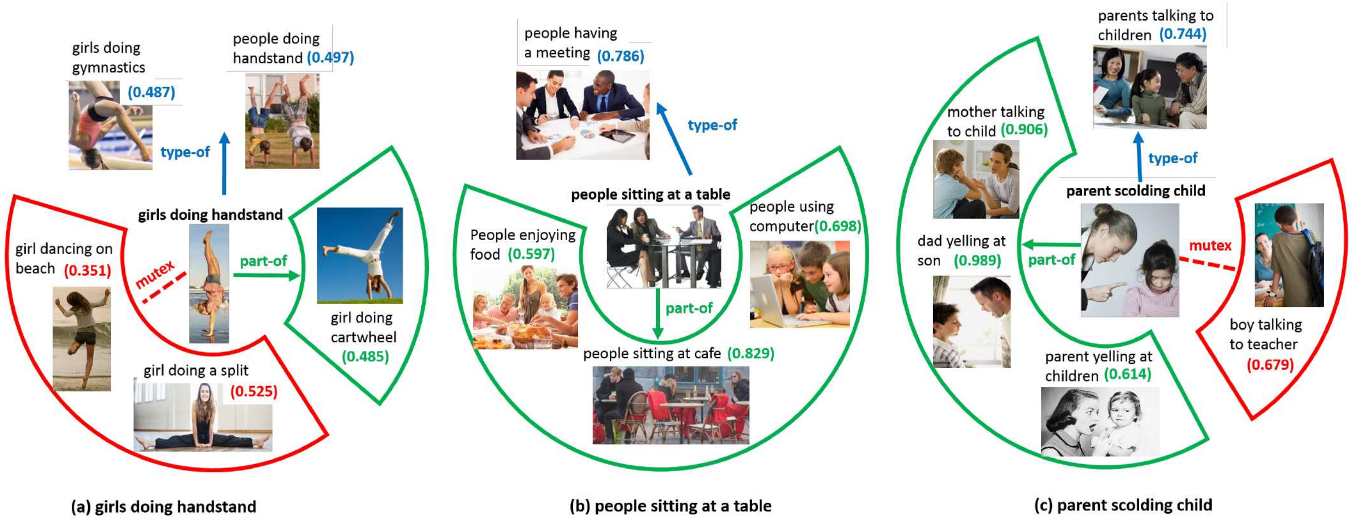


Figure 7. Sample actions where our model achieves more than 10% mAP improvement over RankLoss. The related actions along with relation prediction scores are shown for each of the three actions. Our model effectively treats the images corresponding to the part-of related actions (shown in a green arc) as additional positives, and those of the mutually exclusive actions (shown in a red) as hard negatives.

5. RankLoss This is the basic action retrieval model Sec. 3.2, without the use of relationships.
6. LinearComb The action score of an image is determined by a linear combination of the scores of related actions. The weights are determined by the visual similarity between the training images of the two actions. A higher weight is assigned for a higher similarity. Note that this method is similar to the re-scoring approach from NEIL [2].
7. DeViSE [12] The action embedding layer of Sec. 3.2 is replaced by a fixed embedding vector, which is obtained as the average of the word-vector embeddings of the words present in the action description.
8. ProjectedDeViSE [12] We learn a linear layer on top of the word vector embeddings, similar to [18].
9. OurOnlyLangRel Only Language prior is used to determine relations in our model.
10. OurWeightedLangRel We use from language prior, but we weight the contribution of each related action in our overall objective. The weight is determined by the visual overlap between the two actions. This has the advantage of removing noisy relations.
11. OurNoConsistency Our model without the consistency objective.
12. OurFullModel This is our full model with consistency objective.

The results for the $27K$ and $2K$ action datasets are shown in Tab. 1. The Ranking Loss model outperforms both the Softmax and Logistic models. Since the HEX-graph method provides a generalization of the logistic and softmax models, its performance is also seen to suffer in comparison to RankLoss. We also observe that the performance of the DeViSE model is not significantly better than RankLoss. For composite descriptions like actions, a simple word vector averaging is not seen to capture the visual relationship between the actions. Similarly, a naive use of the language prior is seen to hurt performance in OurOnlyLangRel. Also, a direct fusion of the scores in LinearComb, similar to the approach by NEIL only provides a marginal gain.

| Method | mAP (%) 27K | mAP (%) 2K |
|---|---|---|
| RandomChance | 2.22 | 3.02 |
| Logistic | 5.80 | 5.53 |
| Softmax | 5.79 | 5.47 |
| LangRelWithHEX [8] | 6.01 | 5.96 |
| OurRelWithHEX [8] | 6.43 | 5.71 |
| RankLoss | 8.17 | 6.88 |
| DeViSE [12] | 7.02 | 5.73 |
| ProjectedDeViSE [12] | 7.88 | 6.67 |
| LinearComb | 8.64 | 7.78 |
| OurOnlyLangRel | 6.14 | 7.02 |
| OurWeightedLangRel | 10.05 | 8.87 |
| OurNoConsistency | 11.92 | 10.04 |
| OurFullModel | **11.96** | **10.16** |

Table 1. Results of action retrieval on the 27K and 2K dataset.

Our full model significantly outperforms the previous baselines for both settings. It is also interesting to note that the consistency objective offers only a small advantage in terms of performance, compared to the visual objective in Eq. 8. We visualize a few examples where our model achieves a significant gain compared to RankLoss in Fig. 7. Our performance gain can be attributed to the additional labels extrapolated from the learned relations. In the first example, we see that the action "girl doing a handstand" is identified as part-of "girl doing a


Figure 8. Each row corresponds to an action with a sample test image shown in the first column. Green boxes indicates the test cases, where our model correctly ranked the image higher than RANKLOSS, and the red boxes indicate a lower ranking. The last three columns in each row depict the related actions arranged by decreasing order of relations scores. Correct relation predictions are shown in green, and wrong ones in red.

cartwheel". Hence, the relationship objective in Eq. 8, treats the cartwheel images as additional positives while training a model for handstand. Similarly, by identifying the mutual exclusivity with "girl doing a split", our method gains additional negatives. Since we identify relationships with only those actions which have some overlap in the images returned by image search, a correct mutual exclusion effectively adds hard negatives for training.

**Performance gain from each relationship** We study the impact of each of the three relations on our performance improvement in Fig. 9. For an action, the strength of a specific relation is determined by the sum of the corresponding relation scores with respect to all related actions. At different values of the relation strength, we plot the average improvement in AP of all actions whose corresponding relation strengths are higher than that value. The relationship strength is quantized into 100 bins. We observe that actions which are part-of more actions tend to have the highest improvement in AP, followed by mutual exclusion and type-of. As shown in Fig. 7, we obtain additional positive training data from part-of actions, and negatives form mutually exclusive actions. As a result, we expect these two relations to have a higher impact. This intuition agrees with the plot.

**Evaluating predicted relations** We also present a quantitative evaluation of the predicted relations for a set of 900 action pairs. To clearly see the advantage our method over the naive use of language based relations, we chose those action pairs which do not have a language prior. Further, the action pairs were chosen so that they had an almost unambiguous relationship. The pairs correspond to 1800 (since relationship is asymmetric) relationships. The mean AP of the relationship predictions are shown in Tab. 2. We notice a significant gain in predicting part-of and type-of relations,

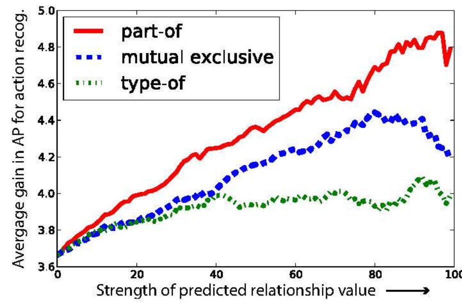


Figure 9. For all three relations, the relation strength for an action is computed as the sum of the corresponding relation scores with respect to its related actions. At each relation strength, we have plotted the average gain (over RANKLOSS) in AP of actions having a relation strength higher than that value. This plot shows that the performance gain is higher for actions with more part-of relations, followed by mutually exclusive and type-of relations.

| Method | mAP(%) for relationship prediction | | |
|---|---|---|---|
| | part-of | type-of | mut-ex |
| RANDOMCHANCE | 36.61 | 36.61 | 34.56 |
| OURFULLMODEL | **60.12** | **60.61** | 42.30 |

Table 2. Results for action relationship prediction for a subset of 900 action pairs (1800 relations).

compared to random chance. This shows the advantage of our method over using only language relations.

**Limitations** While our model provides a significant gain by assigning one of three relationships between action pairs, there are a few instances where the relationship is ambiguous (as shown in failure cases of Fig. 8). Since our model makes soft assignments, these cases can still be partially handled. However, few action pairs have a good visual overlap and an ambiguous relationship such as: "kids doing homework" and "students doing math". Assigning mutual exclusion is seen to hurt performance for these actions.

## 5. Conclusion

In this work, we tackled the problem of learning action retrieval models in a practical setting with a large number of actions which are related to each other. Existing methods achieve a performance gain in such settings by utilizing readily available semantic graphs such as WordNet. However, human actions do not have a predefined semantic graph. We presented a neural network architecture which jointly extracts the relationships between actions and jointly learns better models by extrapolating action labels based on these relations. Our model integrated language cues, visual cues and logical consistency to determine these action relationships. Our full model achieved significant improvement in action retrieval performance over state-of-the-art method [8] for a novel large scale action dataset.

## References


[1] T. Chen, F. X. Yu, J. Chen, Y. Cui, Y.-Y. Chen, and S.-F. Chang. Object-based visual sentiment concept analysis and application. In *Proceedings of the ACM International Conference on Multimedia*, pages 367–376. ACM, 2014.

[2] X. Chen, A. Shrivastava, and A. Gupta. Neil: Extracting visual knowledge from web data. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1409–1416. IEEE, 2013.

[3] J. Choi, M. Rastegari, A. Farhadi, and L. S. Davis. Adding unlabeled samples to categories by learned attributes. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 875–882. IEEE, 2013.

[4] I. Dagan, O. Glickman, and B. Magnini. The pascal recognising textual entailment challenge. In *Machine learning challenges.*, pages 177–190. Springer, 2006.

[5] 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.

[6] T. Dean, M. A. Ruzon, M. Segal, J. Shlens, S. Vijayanarasimhan, and J. Yagnik. Fast, accurate detection of 100,000 object classes on a single machine. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1814–1821. IEEE, 2013.

[7] O. Dekel, J. Keshet, and Y. Singer. Large margin hierarchical classification. In *Proceedings of the twenty-first international conference on Machine learning*, page 27. ACM, 2004.

[8] 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.

[9] J. Deng, J. Krause, A. Berg, and L. Fei-Fei. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, RI, USA, June 2012.

[10] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[11] R. Fergus, H. Bernal, Y. Weiss, and A. Torralba. Semantic label sharing for learning with many categories. In *Computer Vision–ECCV 2010*, pages 762–775. Springer, 2010.

[12] 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.

[13] 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.

[14] G. Guo and A. Lai. A survey on still image based human action recognition. *Pattern Recognition*, 47(10):3343 – 3361, 2014.

[15] J.-F. Hu, W.-S. Zheng, J. Lai, S. Gong, and T. Xiang. Recognising human-object interaction via exemplar based modelling. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3144–3151. IEEE, 2013.

[16] S. J. Hwang, F. Sha, and K. Grauman. Sharing features between objects and their attributes. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1761–1768. IEEE, 2011.

[17] Y. Jia, J. T. Abbott, J. Austerweil, T. Griffiths, and T. Darrell. Visual concept learning: Combining machine vision and bayesian generalization on concept hierarchies. In *Advances in Neural Information Processing Systems*, pages 1842–1850, 2013.

[18] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image senetence mapping. In *Advances in Neural Information Processing Systems*, 2014.

[19] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[21] D. Kuettel, M. Guillaumin, and V. Ferrari. Segmentation propagation in imagenet. In *Computer Vision–ECCV 2012*, pages 459–473. Springer, 2012.

[22] C. Li, D. Parikh, and T. Chen. Automatic discovery of groups of objects for scene understanding. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2735–2742. IEEE, 2012.

[23] J. J. Lim, R. Salakhutdinov, and A. Torralba. Transfer learning by borrowing examples for multiclass object detection. In *Neural Information Processing Systems (NIPS)*, 2011.

[24] B. MacCartney and C. D. Manning. Natural logic for textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 193–200. Association for Computational Linguistics, 2007.

[25] M. Marszalek and C. Schmid. Semantic hierarchies for visual object recognition. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–7. IEEE, 2007.

[26] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[27] V. Ordonez, J. Deng, Y. Choi, A. C. Berg, and T. L. Berg. From large scale image categorization to entry-level categories. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2768–2775. IEEE, 2013.

[28] A. Prest, C. Schmid, and V. Ferrari. Weakly supervised learning of interactions between humans and objects. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(3):601–614, March 2012.

[29] J. Pu, Y.-G. Jiang, J. Wang, and X. Xue. Which looks like which: Exploring inter-class relationships in fine-grained vi-

sual categorization. In *Computer Vision–ECCV 2014*, pages 425–440. Springer, 2014.

[30] V. Ramanathan, P. Liang, and L. Fei-Fei. Video event understanding using natural language descriptions. In *International Conference on Computer Vision (ICCV)*, 2013.

[31] R. Salakhutdinov, A. Torralba, and J. Tenenbaum. Learning to share visual appearance for multiclass object detection. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1481–1488. IEEE, 2011.

[32] G. Sharma, F. Jurie, and C. Schmid. Discriminative spatial saliency for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, June 2012.

[33] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199*, 2014.

[34] R. Socher, D. Chen, C. D. Manning, and A. Y. Ng. Reasoning with neural tensor networks for knowledge base completion. In *Advances in Neural Information Processing Systems*, 2013.

[35] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *TACL*, 2013.

[36] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[37] N. Verma, D. Mahajan, S. Sellamanickam, and V. Nair. Learning hierarchical similarity metrics. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2280–2287. IEEE, 2012.

[38] G. Wang, D. Forsyth, and D. Hoiem. Improved object categorization and detection using comparative object similarity. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(10):2442–2453, 2013.

[39] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 103(1):60–79, 2013.

[40] J. Wang, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, Y. Wu, et al. Learning fine-grained image similarity with deep ranking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[41] B. Yao and L. Fei-Fei. Grouplet: A structured image representation for recognizing human and object interactions. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 9–16. IEEE, 2010.

[42] B. Yao and L. Fei-Fei. Action recognition with exemplar based 2.5 d graph matching. In *Computer Vision–ECCV 2012*, pages 173–186. Springer, 2012.

[43] 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.

[44] B. Zhao, F. Li, and E. P. Xing. Large-scale category structure aware image categorization. In *Advances in Neural Information Processing Systems*, pages 1251–1259, 2011.

[45] X. Zhu, D. Anguelov, and D. Ramanan. Capturing long-tail distributions of object subcategories. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2014.

[46] Y. Zhu, A. Fathi, and L. Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *Computer Vision–ECCV 2014*, pages 408–424. Springer, 2014.

[47] A. Zweig and D. Weinshall. Exploiting object hierarchy: Combining models from different category levels. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.