Compositional Learning for Human Object Interaction

Keizo Kato[1(✉)], Yin Li[2], and Abhinav Gupta[2]

[1] Fujitsu Laboratories Ltd., Kawasaki, Japan
kato.keizo@jp.fujitsu.com
[2] Carnegie Mellon University, Pittsburgh, USA
yinl2@andrew.cmu.edu, abhinavg@cs.cmu.edu

Abstract. The world of human-object interactions is rich. While generally we sit on chairs and sofas, if need be we can even sit on TVs or top of shelves. In recent years, there has been progress in modeling actions and human-object interactions. However, most of these approaches require lots of data. It is not clear if the learned representations of actions are generalizable to new categories. In this paper, we explore the problem of zero-shot learning of human-object interactions. Given limited verb-noun interactions in training data, we want to learn a model than can work even on unseen combinations. To deal with this problem, In this paper, we propose a novel method using external knowledge graph and graph convolutional networks which learns how to compose classifiers for verb-noun pairs. We also provide benchmarks on several dataset for zero-shot learning including both image and video. We hope our method, dataset and baselines will facilitate future research in this direction.

1 Introduction

Our daily actions and activities are rich and complex. Consider the examples in Fig. 1(a). The same verb "sit" is combined with different nouns (chair, bed, floor) to describe visually distinctive actions ("sit on chair" vs. "sit on floor"). Similarly, we can interact with the same object (TV) in many different ways (turn on, clean, watch). Even small sets of common verbs and nouns will create a huge combination of action labels. It is highly unlikely that we can capture action samples covering all these combinations. What if we want to recognize an action category that we had never seen before, e.g., the one in Fig. 1(b)?

This problem is known as zero shot learning, where categories at testing time are not presented during training. It has been widely explored for object recognition [1,11,12,15,31,37,60]. And there is an emerging interest for zero-shot action recognition [18,21,24,35,51,55]. How are actions different from objects in zero shot learning? What we know is that human actions are naturally compositional and humans have amazing ability to achieve similar goals with different objects and tools. For example, while one can use hammer for the hitting the nail, we can

Work was done when K. Kato was at CMU.

V. Ferrari et al. (Eds.): ECCV 2018, LNCS 11218, pp. 247–264, 2018.
https://doi.org/10.1007/978-3-030-01264-9_15

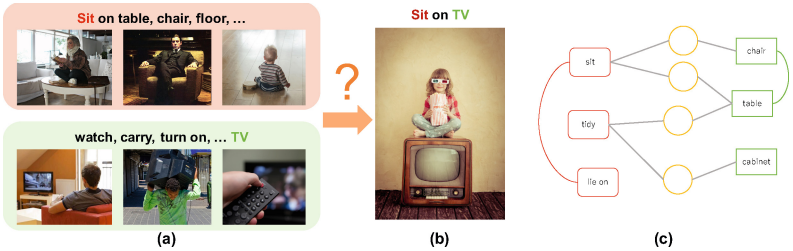

Fig. 1. (a–b) many of our daily actions are compositional. These actions can be described by motion (verbs) and the objects (nouns). We build on this composition for zero shot recognition of human-object interactions. Our method encodes motion and object cues as visual embeddings of verbs (e.g., sit) and nouns (e.g., TV), uses external knowledge for learning to assemble these embeddings into actions. We demonstrate that our method can generalize to unseen action categories (e.g., sit on a TV). (c) a graph representation of interactions: pairs of verb-noun nodes are linked via action nodes (circle), and verb-verb/noun-noun pairs can be connected.

also use a hard-cover book for the same. We can thus leverage this unique composition to help recognizing novel actions. To this end, we address the problem of zero shot action recognition. And we specifically focus on the compositional learning of daily human object interactions, which can be described by a pair of verb and noun (e.g., "wash a mirror" or "hold a laptop").

This compositional learning faces a major question: How can a model learn to compose a novel action within the context? For example, "Sitting on a TV" looks very different from "Sitting on a chair" since the underlying body motion and body poses are quite different. Even if the model has learned to recognize individual concepts like "TV" and "Sitting", it will still fail to generalize. Indeed, many of our seemly effortless interactions with novel objects build on our prior knowledge. If the model knows that people also sit on floor, vase are put on floor, and vase can be put on TV. It might be able to assemble the visual concepts of "Sitting" and "TV" to recognize the rare action of "Sitting on a TV". Moreover, what if model knows "sitting" is similar to "lean" and "TV" is similar to "Jukebox", can model also recognize "lean into Jukebox"? Thus, we propose to explore using external knowledge to bridge the gap of contextuality, and to help the modeling of compositionality for human object interactions.

Specifically, we extract Subject, Verb and Object (SVO) triplets from knowledge bases [8,30] to build an external knowledge graph. These triplets capture a large range of human object interactions, and encode our knowledge about actions. Each verb (motion) or noun (object) is a node in the graph with its word embedding as the node's feature. Each SVO-triplet defines an action node and a path between the corresponding verb and noun nodes via the action node (See Fig. 1(c)). These action nodes start with all zero features, and must learn its representation by propagating information along the graph during training. This information passing is achieved by using a multi-layer graph convolutional

network [29]. Our method jointly trains a projection of visual features and the graph convolutional network, and thus learns to transform both visual features and action nodes into a shared embedding space. Our zero shot recognition of actions is thus reduced to nearest neighbor search in this space.

We present a comprehensive evaluation of our method on image datasets (HICO [7] and a subset of Visual Genome [30]), as well as a more challenging *video* dataset (Charades [48]). We define proper benchmarks for zero shot learning of human-object interactions, and compare our results to a set of baselines. Our method demonstrates strong results for unseen combinations of known concepts. Our results outperforms the state-of-the-art methods on HICO and Visual Genome, and performs comparably to previous methods on Charades. We also show that our method can generalize to unseen concepts, with a performance level that is much better than chance. We hope our method and benchmark will facilitate future research in this direction.

2 Related Work

Zero Shot Learning. Our work follows the zero-shot learning setting [53]. Early works focused on attribute based learning [26,31,41,58]. These methods follow a two-stage approach by first predicting the attributes, and then inferring the class labels. Recent works make use of semantic embeddings to model relationships between different categories. These methods learn to map either visual features [15,55], or labels [1,11,12,37], or both of them [52,52,56] into a common semantic space. Recognition is then achieved by measuring the distance between the visual inputs and the labels in this space. Similar to attribute based approaches, our method considers interactions as verb-noun pairs. However, we do not explicit predict individual verbs or nouns. Similar to embedding based approaches, we learn semantic embeddings of interactions. Yet we focus on the compositional learning [40] by leveraging external knowledge.

Our work is also related to previous works that combine side information for zero shot recognition. For example, Rohrbach et al. [43] transferred part attributes from linguistic data to recognize unseen objects. Fu et al. [16] used hyper-graph label propagation to fuse information from multiple semantic representations. Li et al. [33] explored semi-supervised learning in a zero shot setting. Inspired by these methods, our method connects actions and objects using information from external knowledge base. Yet we use graph convolution to propagate the semantic representations of verbs and nouns, and learns to assemble them into actions. Moreover, previous works considered the recognition of objects in images. Our work thus stands out by addressing the recognition of human object interactions in both images and videos. We believe our problem is an ideal benchmark for compositional learning of how to build generalizable representations.

Modeling Human Object Interactions. Modeling human object interactions has a rich history in both computer vision and psychology. It starts from the idea of "affordances" introduced by Gibson [17]. There have been lots of work in using semantics for functional understanding of objects [49]. However, none

of these early attempts scaled up due to lack of data and brittle inference under noisy perception. Recently, the idea of modeling human object interactions has made a comeback [19]. Several approaches have looked at modeling semantic relationships [10,20,57], action-3D relationships [14] or completely data-driven approach [13]. However, none of them considered the use of external knowledge.

Moreover, recent works focused on creating large scale image datasets for human object interactions [7,30,36]. However, even the current largest dataset—Visual Genome [30] only contains a small subset of our daily interactions (hundreds), and did not capture the full dynamics of interactions that exist in video. Our work takes a step forward by using external knowledge for recognizing unseen interactions, and exploring the recognition of interactions for a challenging video dataset [48]. We believe an important test of intelligence and reasoning is the ability to compose primitives into novel concepts. Therefore, we hope our work can provide a step for visual reasoning based approaches to come in future.

Zero Shot Action Recognition. Our paper is inspired by compositional representations for human object interactions. There has been a lot of work in psychology and early computer vision on compositions, starting from original work by Biederman [4] and Hoffman et al. [23]. More recently, several works started to address the zero shot recognition of actions. Similar to attribute based object recognition, Liu et al. [35] learned to recognize novel actions using attributes. Going beyond recognition, Habibian et al. [21] proposed to model concepts in videos for event detection. Inspired by zero shot object recognition, Xu et al. presented a embedding based method for actions [55]. Other efforts include the exploration of text descriptions [18,51], joint segmentation of actors and actions [54], and model domain shift of actions [56]. However, these methods simply treat actions as labels and did not consider their compositionality.

Perhaps the most relevant work is from [24,25,28]. Jain et al. [24,25] noticed a strong relation between objects and actions, and thus proposed to use object classifier for zero shot action recognition. As a step forward, Kalogeition et al. [28] proposed to jointly detect objects and actions in videos. Instead of using objects alone, our method models both body motion (verb) and objects (noun). More importantly, we explore using external knowledge for assembling these concepts into novel actions. Our method thus provides a revisit to the problem of human object interactions from the perspective of compositionality.

Compositional Learning for Vision and Language. Compositional learning has been explored in Visual Question Answering (VQA). Andreas et al. [2,3] decomposed VQA task into sequence of modular sub-problems—each modeled by a neural network. Their method assembles a network from individual modules based on the syntax of a question, and predicts the answer using the instance-specific network. This idea was further extended by Johnson et al. [27], where deep models are learned to generate programs from a question and to execute the programs on the image to predict the answer. Our method shares the core idea of compositional learning, yet focuses on human object interactions. Moreover, modeling SVO pairs using graph representations has been discussed in [45,50,59]. Sadeghi et al. [45] constructed a knowledge graph of SVO nodes

similar to our graph representation. However, their method aimed at verifying SVO relationships using visual data. A factor graph model with SVO nodes was presented in for video captioning [50], yet without using deep models. More recently, Zellers et al. [59] proposed a deep model for generating scene graphs of objects and their relations from an image. However, their method can not handle unseen concepts.

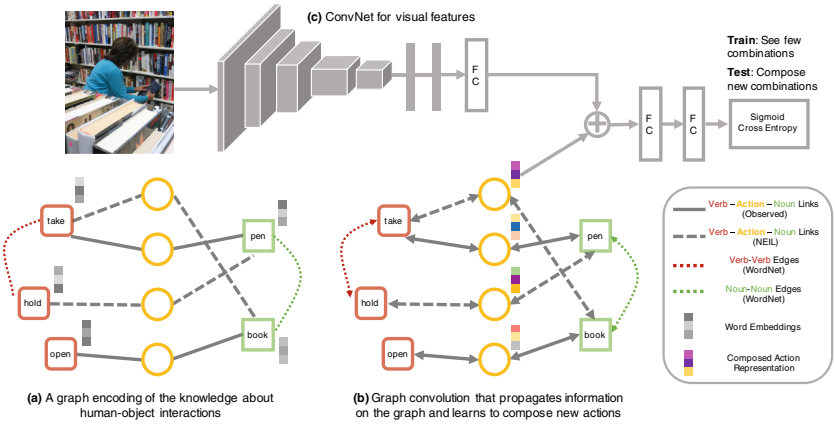

Fig. 2. Overview of our approach. (a) our graph that encodes external SVO pairs. Each verb or noun is represented as a node and comes with its word embeddings as the node's features. Every interaction defined by a SVO pair creates a new action node (orange ones) on the graph, which is linked to the corresponding noun and verb nodes. We can also add links between verbs and nouns, e.g., using WordNet [39]. (b) the graph convolution operation. Our learning will propagate features on the graph, and fill in new representations for the action nodes. These action features are further merged with visual features from a convolutional network (c) to learn a similarity metric between the action concepts and the visual inputs. (Color figure online)

3 Method

Given an input image or video, we denote its visual features as x_i and its action label as y_i. We focus on human object interactions, where y_i can be further decomposed into a verb y_i^v (e.g., "take"/"open") and a noun y_i^n (e.g., "phone"/"table"). For clarity, we drop the subscript i when it is clear that we refer to a single image or video. In our work, we use visual features from convolutional networks for x, and represent verbs y^v and nouns y^n by their word embeddings as z^v and z^n.

Our goal is to explore the use of knowledge for zero shot action recognition. Specifically, we propose to learn a score function ϕ such that

$$p(y|x) = \phi(x, y^v, y^n; \mathcal{K}) \tag{1}$$

where $\mathcal{K}$ is the prior knowledge about actions. Our key idea is to represent $\mathcal{K}$ via a graph structure and use this graph for learning to compose representations of novel actions. An overview of our method is shown in Fig. 2. The core component of our model is a graph convolutional network $g(y^v, y^n; \mathcal{K})$ (See Fig. 2(a–b)). g learns to compose action representation z_a based on embeddings of verbs and nouns, as well as the knowledge of SVO triplets and lexical information. The output z_a is further compared to the visual feature x for zero shot recognition. We now describe how we encode external knowledge using a graph, and how we use this graph for compositional learning.

3.1 A Graphical Representation of Knowledge

Formally, we define our graph as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, Z)$. $\mathcal{G}$ is a undirected graph with $\mathcal{V}$ as its nodes. $\mathcal{E}$ presents the links between nodes $\mathcal{V}$ and Z are the feature vectors for nodes $\mathcal{E}$. We propose to use this graph structure to encode two important types of knowledge: (1) the "affordance" of objects, such as "book can be hold" or "pen can be taken", defined by SVO triplets from external knowledge base [8]; (2) the semantic similarity between verb or noun tokens, defined by the lexical information from WordNet [39].

Graph Construction. Specifically, we construct the graph as follows.

- Each verb or noun is modeled as a node on the graph. These nodes are denoted as $\mathcal{V}_v$ and $\mathcal{V}_n$. And they comes with their word embeddings [38,42] as the nodes features Z_v and Z_n
- Each verb-object pair in a SVO defines a human object interaction. These interactions are modeled by a separate set of action nodes $\mathcal{V}_a$ on the graph. Each interaction will have its own node, even if it share the same verb or noun with other interactions. For example, "take a book" and "hold a book" will be two different nodes. These nodes are initialized with all zero feature vectors, and must obtain their representation Z_a via learning.
- A verb node can only connect to a noun node via a valid action node. Namely, each interaction will add a new path on the graph.
- We also add links within noun or verb nodes by WordNet [39].

This graph is thus captured by its adjacency matrix $\mathcal{A} \in R^{|\mathcal{V}|\times|\mathcal{V}|}$ and a feature matrix $Z \in R^{d\times|\mathcal{V}|}$. Based on the construction, our graph structure can be naturally decomposed into blocks, given by

$$\mathcal{A} = \begin{bmatrix} \mathcal{A}_{vv} & 0 & \mathcal{A}_{va} \\ 0 & \mathcal{A}_{nn} & \mathcal{A}_{an}^T \\ \mathcal{A}_{va}^T & \mathcal{A}_{an} & 0 \end{bmatrix}, \quad Z = [Z_v, Z_n, 0] \tag{2}$$

where $\mathcal{A}_{vv}$, $\mathcal{A}_{va}$, $\mathcal{A}_{an}$, $\mathcal{A}_{nn}$ are adjacency matrix for verb-verb pairs, verb-action pairs, action-noun pairs and noun-noun pairs, respectively. Z_v and Z_n are word embedding for verbs and nouns. Moreover, we have $Z_a = 0$ and thus the action nodes need to learn new representations for recognition.

Graph Normalization. To better capture the graph structure, it is usually desirable to normalize the adjacency matrix [29]. Due to the block structure in our adjacency matrix, we add an identity matrix to the diagonal of $\mathcal{A}$, and normalize each block separately. More precisely, we have

$$\hat{\mathcal{A}} = \begin{bmatrix} \hat{\mathcal{A}}_{vv} & 0 & \hat{\mathcal{A}}_{va} \\ 0 & \hat{\mathcal{A}}_{nn} & \hat{\mathcal{A}}_{an}^T \\ \hat{\mathcal{A}}_{an} & \hat{\mathcal{A}}_{va}^T & I \end{bmatrix}, \tag{3}$$

where $\hat{\mathcal{A}}_{vv} = D_{vv}^{-\frac{1}{2}}(\mathcal{A}_{vv}+I)D_{vv}^{\frac{1}{2}}$, $\hat{\mathcal{A}}_{nn} = D_{nn}^{-\frac{1}{2}}(\mathcal{A}_{nn}+I)D_{nn}^{\frac{1}{2}}$, $\hat{\mathcal{A}}_{va} = D_{va}^{-\frac{1}{2}}\mathcal{A}_{vv}D_{va}^{\frac{1}{2}}$ and $\hat{\mathcal{A}}_{vn} = D_{vn}^{-\frac{1}{2}}(\mathcal{A}_{vn}+I)D_{vn}^{\frac{1}{2}}$. D is the diagonal node degree matrix for each block. Thus, these are symmetric normalized adjacency blocks.

3.2 Graph Convolutional Network for Compositional Learning

Given the knowledge graph $\mathcal{G}$, we want to learn to compose representation of actions Z_a. Z_a can thus be further used as "action template" for zero shot recognition. The question is how can we leverage the graph structure for learning Z_a. Our key insight is that word embedding of verbs and nouns encode important semantic information, and we can use the graph to distill theses semantics, and construct meaningful action representation. To this end, we adopt the Graph Convolution Network (GCN) from [29]. The core idea of GCN is to transform the node features based on its neighbors on the graph. Formally, given normalized graph adjacency matrix $\hat{\mathcal{A}}$ and node features Z, a single layer GCN is given by

$$\tilde{Z} = GCN(Z, A) = \hat{\mathcal{A}}Z^T W \tag{4}$$

where W is a $d \times \tilde{d}$ weight learned from data. d is the dimension of input feature vector for each node and $\tilde{d}$ is the output feature dimension. Intuitively, GCN first transforms each feature on each node independently, then averages the features of connected nodes. This operation is usually stacked multiple times, with nonlinear activation functions (ReLU) in-between.

Note that $\hat{\mathcal{A}}$ is a block matrix. It is thus possible to further decompose GCU operations to each block. This decomposition provides better insights to our model, and can significantly reduce the computational cost. Specially, we have

$$\tilde{Z}_v = \hat{\mathcal{A}}_{vv}Z_v^T W_{vv} \quad \tilde{Z}_n = \hat{\mathcal{A}}_{nn}Z_n^T W_{nn} \quad \tilde{Z}_a = \hat{\mathcal{A}}_{an}Z_v^T W_{an} + \mathcal{A}_{va}^T Z_n^T W_{va} \tag{5}$$

where $W_{vv} = W_{nn} = W_{an} = W_{va} = W$. We also experimented with using different parameters for each block, which is similar to [46]. Note the last line of $\tilde{Z}_a$ in Eq. 5. In a single layer GCN, this model learns linear functions W_{an} and W_{va} that transform the neighboring word embeddings into an action template. With nonlinear activations and K GCN layers, the model will construct a nonlinear transform that considers more nodes for building the action representation (from 1-neighborhood to K-neighborhood).

3.3 From Graph to Zero Shot Recognition

The outputs of our graph convolutional networks are the transformed node features $\tilde{Z} = [\tilde{Z}_v, \tilde{Z}_n, \tilde{Z}_a]$. We use the output action representations $\tilde{Z}_a$ for the zero shot recognition. This is done by learning to match action features $\tilde{Z}_a$ and visual features x. More precisely, we learn a score function h that takes the inputs of $\tilde{Z}_a$ and x, and outputs a similarity score between $[0, 1]$.

$$h(x, a) = h(f(x) \oplus \tilde{Z}_a) \tag{6}$$

where f is a nonlinear transform that maps x into the same dimension as $\tilde{Z}_a$. $\oplus$ denotes concatenation. h is realized by a two-layer network with sigmoid function at the end. h can be considered as a variant of a Siamese network [9].

3.4 Network Architecture and Training

We present the details about our network architecture and our training.

Architecture. Our network architecture is illustrated in Fig. 2. Specifically, our model includes 2 graph convolutional layers for learning action representations. Their output channels are 512 and 200, with ReLU units after each layer. The output of GCN is concatenated with image features from a convolutional network. The image feature has a reduced dimension of 512 by a learned linear transform. The concatenated feature vector as sent to two Fully Connected (FC) layer with the size of 512 and 200, and finally outputs a scalar score. For all FC layers except the last one, we attach ReLU and Dropout (ratio = 0.5).

Training the Network. Our model is trained with a logistic loss attached to g. We fix the image features, yet update all parameters in GCN. We use mini-batch SGD for the optimization. Note that there are way more negative samples (unmatched actions) than positive samples in a mini-batch. We re-sample the positives and negatives to keep the their ratio fixed (1:3). This re-sampling strategy prevents the gradients to be dominated by the negative samples, and thus is helpful for learning. We also experimented with hard-negative sampling, yet found that it leads to severe overfitting on smaller datasets.

4 Experiments

We now present our experiments and results. We first introduce our experiment setup, followed by a description of the datasets and baselines. Finally, we report our results and compare them to state-of-the-art methods.

4.1 Experiment Setup

Benchmark. Our goal is to evaluate if methods can generalize to unseen actions. Given the compositional structure of human-object interactions, these unseen actions can be characterized into two settings: (a) a novel combination of known

noun and verb; and (b) a new action with unknown verbs or nouns or both of them. We design two tasks to capture both settings. Specifically, we split both noun and verb tokens into two even parts. We denote the splits of nouns as 1/2 and verbs as A/B. Thus, 1B refers to actions from the first split of nouns and the second split of verbs. We select combinations of the splits for training and testing as our two benchmark tasks.

- **Task 1.** Our first setting allows a method to access the full set of verbs and nouns during training, yet requires the method to recognize either a seen or an unseen combination of known concepts for testing. For example, a method is given the action of "hold apple" and "wash motorcycle", and is asked to recognize novel combinations of "hold motorcycle" and "wash apple". Our training set is a subset of 1A and 2B (1A + 2B). This set captures all concepts of nouns and verbs, yet misses many combination of them (1B/2A). Our testing set consists of samples from 1A and 2B and unseen combination of 1B and 2A.
- **Task 2.** Our second setting exposes only a partial set of verbs and nouns (1A) to a method during training. But the method is tasked to recognize all possible combinations of actions (1A, 1B, 2A, 2B), including those with unknown concepts. For example, a method is asked to jump from "hold apple" to "hold motorcycle" and "wash apple", as well as the complete novel combination of "wash motorcycle". This task is extremely challenging. It requires the method to generalize to completely new categories of nouns and verbs, and assemble them into new actions. We believe the prior knowledge such as word-embeddings or SVO pairs will allow the jumps from 1 to 2 and A to B. Finally, we believe this setting provides a good testbed for knowledge representation and transfer.

Generalized Zero Shot Learning. We want to highlight that our benchmark follows the setting of generalized zero shot learning [53]. Namely, during test, we did no constrain the recognition to the categories on the test set but all possible categories. For example, if we train on 1A, during testing the output class can be any of $\{1A, 2B, 2A, 2B\}$. We do also report numbers separately for each subset to understand where what approach works. More importantly, as pointed out by [53], a ImageNet pre-trained model may bias the results if the categories are already seen during pre-training. We force nouns that appears in ImageNet [44] stay in training sets for all our experiments except for Charades.

Mining from Knowledge Bases. We describe how we construct the knowledge graph for all our experiments. Specifically, we make use of WordNet to create noun-noun and verb-verb links. We consider two nodes are connected if (1) they are the immediate hypernym or hyponym to each other (denoted as 1 HOP); (2) their LCH similarity score [32] is larger than 2.0. Furthermore, we extracted SVO from NELL [5] and further verified them using COCO dataset [34]. Specifically, we parse all image captions on COCO, only keep the verb-noun pairs that appeared on COCO, and add the remaining pairs to our graph.

Implementation Details. We extracted the last FC features from ResNet 152 [22] pre-trained with ImageNet for HICO and Visual Genome HOI datasets, and I3D Network pre-trained with kinetics [6] for Charades dataset. All images are re-sized to 224×224 and the convolutional network is fixed. For all our experiments, we used GloVe [42] for embedding verb and noun tokens, leading to a 200D vector for each token. GloVe is pretrained with Wikipedia and Gigaword5 text corpus. We adapt hard negative mining for HICO and Visual Genome HOI datasets, yet disable it for Charades dataset to prevent overfitting.

Table 1. Ablation study of our methods. We report mAP for both tasks and compare different variant of our methods. These results suggest that adding more links to the graph (and thus inject more prior knowledge) helps to improve the results.

Methods	mAP on test set			
	Train 1A + 2B		Train 1A	
	All	2A + 1B Unseen	All	1B + 2A + 2B Unseen
Chance	0.55	0.49	0.55	0.51
GCNCL-I	20.96	16.05	11.93	7.22
GCNCL-I + A	21.39	16.82	11.57	6.73
GCNCL-I + NV + A	**21.40**	**16.99**	11.51	6.92
GCNCL	19.91	14.07	11.46	7.18
GCNCL + A	20.43	15.65	11.72	7.19
GCNCL + NV + A	21.04	16.35	**11.94**	**7.50**

4.2 Dataset and Benchmark

We evaluate our method on HICO [7], Visual Genome [30] and Charades [48] datasets. We use mean Average Precision (mAP) scores averaged across all categories as our evaluation metric. We report results for both tasks (unseen combination and unseen concepts). We use 80/20 training/testing splits for all experiments unless otherwise noticed. Details of these datasets are described below.

HICO Dataset [7] is developed for Humans Interacting with Common Objects. It is thus particularly suitable for our task. We follow the classification task. The goal is to recognize the interaction in an image, with each interaction consists of a verb-noun pair. HICO has 47,774 images with 80 nouns, 117 verbs and 600 interactions. We remove the verb of "no interaction" and all its associated categories. Thus our benchmark of HICO includes 116 verbs and 520 actions.

Visual Genome HOI Dataset is derived from Visual Genome [30]—the largest dataset for structured image understanding. Based on the annotations, we carve out a sub set from Visual Genome that focuses on human object interactions. We call this dataset Visual Genome HOI in our experiments. Specifically, from all annotations, we extracted relations in the form of "human-verb-object"

and their associated images. Note that we did not include relations with "be", "wear" or "have", as most of these relations did not demonstrate human object interactions. The Visual Genome HOI dataset includes 21256 images with 1422 nouns, 520 verbs and 6643 unique actions. We notice that a large amount of actions only have 1 or 2 instances. Thus, for testing, we constrain our actions to 532 categories, which include more than 10 instances.

Charades Dataset [48] contains 9848 videos clips of daily human-object interactions that can be described by a verb-noun pair. We remove actions with "no-interaction" from the original 157 category. Thus, our benchmark on Charades includes interactions with 37 objects and 34 verbs, leading to a total of 149 valid action categories. We note that Charades is a more challenging dataset as the videos are captured in naturalistic environments.

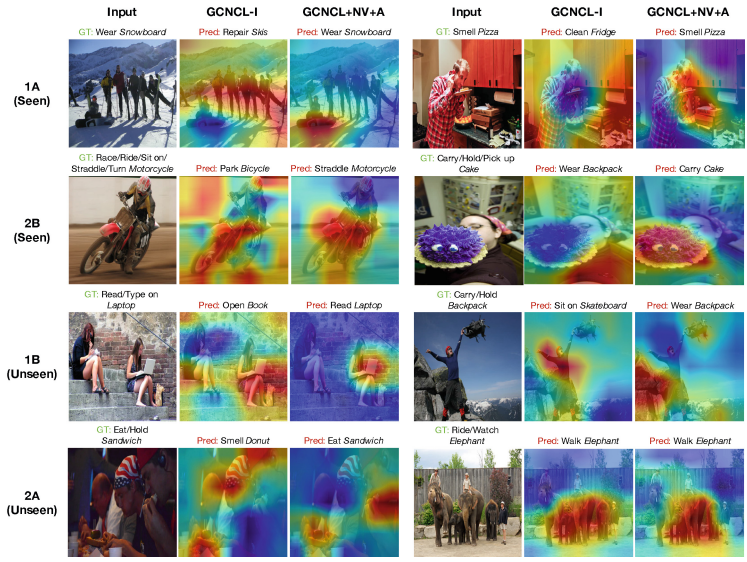

Fig. 3. Results of GCNCL-I and GCNCL + NV + A on HICO dataset. All methods are trained on 1A + 2B and tested on both seen (1A, 2B) and unseen (2A, 1B) actions. Each row shows results on a subset. Each sample includes the input image and its label, top-1 predictions from GCNCL-I and GCNCL + NV + A. We plot the attention map using the top-1 predicted labels. Red regions correspond to high prediction scores. (Color figure online)

4.3 Baseline Methods

We consider a set of baselines for our experiments. These methods include

- **Visual Product** [31] **(VP):** VP composes outputs of a verb and a noun classifier by computing their product ($p(a, b) = p(a)p(b)$). VP does not model

contextuality between verbs and nouns, and thus can be considered as late fusion. VP can deal with unseen combination of known concepts but is not feasible for novel actions with unknown verb or noun.

- **Triplet Siamese Network (Triplet Siamese):** Triplet Siamese is inspired by [12,15]. We first concatenate verb and noun embedding and pass them through two FC layers (512, 200). The output is further concatenated with visual features, followed by another FC layers to output a similarity score. The network is trained with sigmoid cross entropy loss.
- **Semantic Embedding Space (SES)** [55]**:** SES is originally designed for zero shot action recognition. We take the average of verb and noun as the action embedding. The model learns to minimize the distance between the action embeddings and their corresponding visual features using L2 loss.
- **Deep Embedding Model** [60] **(DEM):** DEM passes verb and noun embeddings independently through FC layers. Their outputs are fused (element-wise sum) and matched to visual features using L2 loss.
- **Classifier Composition** [40] **(CC):** CC composes classifiers instead of word embeddings. Each token is represented by its SVM classifier weights. CC thus learns to transform the combination of two weights into the new classifier. The model is trained with sigmoid cross entropy loss. It can not handle novel concepts if no samples are provided for learning the classifier.

4.4 Ablation Study

We start with an ablation study of our method. We denote our base model as GCNCL (Graph Convolutional Network for Compositional Learning) and consider the following variants

- **GCNCL-I** is our base model that only includes action links on the dataset. There is no connection between nouns and verbs in this model and thus the adjacency matrix of $\mathcal{A}_{vv}$ and $\mathcal{A}_{nn}$ are identity matrix.
- **GCNCL** further adds edges within noun/verb nodes using WordNet.
- **GCNCL/GCNCL-I + A** adds action links from external knowledge base.
- **GCNCL/GCNCL-I + NV + A** further includes new tokens (1 Hop on WordNet). Note that we did not add new tokens for Visual Genome dataset.

We evaluate these methods on HICO dataset and summarize the results in Table 1. For recognizing novel combination of seen concepts, GCNCL-I works better than GCNCL versions. We postulate that removing these links will force the network to pass information through action nodes, and thus help better compose action representations from seen concepts. However, when tested with a more challenging case of recognizing novel concepts, the results are in favor of GCNCL model, especially on the unseen categories. In this case, the model has to use the extra links (verb-verb or noun-noun) for learning the represent ions for new verbs and nouns. Moreover, for both settings, adding more links generally helps to improve the performance, independent of the design of the model. This result provides a strong support to our core argument—external knowledge can be used to improve zero shot recognition of human object interactions.

Moreover, we provide qualitative results in Fig. 3. Specifically, we compare the results of GCNCL-I and GCNCL + NV + A and visualize their attention maps using Grad-Cam [47]. Figure 3 helps to understand the benefit of external knowledge. First, adding external knowledge seems to improve the recognition of nouns but not verbs. For example, GCNCL + NV + A successfully corrected the wrongly recognized objects by GCNCL-I (e.g., "bicycle" to "motorcycle", "skateboard" to "backpack"). Second, both methods are better at recognizing nouns—objects in the interactions. And their attention maps highlight the corresponding object regions. Finally, mis-matching of verbs is the main failure mode of our methods. For the rest of our experiments, we only include the best performing methods of GCNCL-I + NV + A and GCNCL + NV + A.

4.5 Results

We present the full results of our methods and compare them to our baselines.

HICO. Our methods outperformed all previous methods when tasked to recognize novel combination of actions. Especially, our results for the unseen categories achieved a relative gap of 6% when compared to the best result from previous work. When tested on more challenging task 2, our results are better overall, yet slightly worse than Triplet Siamese. We further break down the results on different test splits. It turns out that our result is only worse on the split of 1B (−2.8%), where the objects have been seen before. And our results are better in all other cases (+2.0% on 2A and +0.9% on 2B). We argue that Triplet Siamese might have over-fitted to the seen object categories, and thus will fail to transfer knowledge to unseen concepts. Moreover, we also run significance analysis to explore if the results are statistically significant. We did t-test by comparing results of our GCNCL-I + NV + A to CC (training on 1A + 2B) and GCNCL + NV + A to Triplet Siamese (training on 1A) for all classes. Our results are significantly better than CC (P = 0.04) and Triplet Siamese (P = 0.05) (Tables 2 and 3).

Table 2. Recognition results (mAP) on HICO. We benchmark both tasks of recognizing unseen combinations of known concepts and of recognizing novel concepts.

Methods	mAP on test set			
	Train 1A + 2B		Train 1A	
	All	2A + 1B Unseen	All	1B + 2A + 2B Unseen
Chance	0.55	0.49	0.55	0.51
Triplet Siamese	17.61	16.40	10.38	**7.76**
SES	18.39	13.00	11.69	7.19
DEM	12.26	11.33	8.32	6.06
VP	13.96	10.83	-	-
CC	20.92	15.98	-	-
GCNCL-I + NV + A	**21.40**	**16.99**	11.51	6.92
GCNCL + NV + A	21.04	16.35	**11.94**	7.50

Table 3. Results (mAP) on Visual Genome HOI. This is a very challenging dataset with many action classes and few samples per class.

Methods	mAP on test set			
	Train 1A + 2B		Train 1A	
	All	2A + 1B Unseen	All	1B + 2A + 2B Unseen
Chance	0.28	0.25	0.28	0.32
Triplet Siamese	5.68	4.61	2.55	1.67
SES	2.74	1.91	2.07	0.96
DEM	3.82	3.73	2.26	1.5
VP	3.84	2.34	-	-
CC	6.35	**5.74**	-	-
GCNCL-I + A	6.48	5.10	4.00	**2.63**
GCNCL + A	**6.63**	5.42	**4.07**	2.44

Visual Genome. Our model worked the best except for unseen categories on our first task. We note that this dataset is very challenging as there are more action classes than HICO and many of them have only a few instances. We want to highlight our results on task 2, where our results show a relative gap of more than 50% when compared to the best of previous method. These results show that our method has the ability to generalize to completely novel concepts (Table 4).

Table 4. Results (mAP) on Charades dataset. This is our attempt to recognize novel interactions in videos. While the gap is small, our method still works the best.

Methods	mAP on test set			
	Train 1A + 2B		Train 1A	
	ALL	2A + 1B Unseen	ALL	1B + 2A + 2B Unseen
Chance	1.37	1.45	1.37	1.00
Triplet Siamese	14.23	10.1	10.41	7.82
SES	13.12	9.56	10.14	7.81
DEM	11.78	8.97	9.57	7.74
VP	13.66	9.15	-	-
CC	14.31	10.13	-	-
GCNCL-I + A	**14.32**	10.34	10.48	7.95
GCNCL + A	**14.32**	**10.48**	**10.53**	**8.09**

Charades. Finally, we report results on Charades—a video action dataset. This experiment provides our first step towards recognizing realistic interactions in videos. Again, our method worked the best among all baselines. However, the gap is smaller on this dataset. Comparing to image datasets, Charades has less number of samples and thus less diversity. Methods can easily over-fit on this dataset. Moreover, building video representations is still an open challenge. It might be that our performance is limited by the video features.

5 Conclusion

We address the challenging problem of compositional learning of human object interactions. Specifically, we explored using external knowledge for learning to compose novel actions. We proposed a novel graph based model that incorporates knowledge representation into a deep model. To test our method, we designed careful evaluation protocols for zero shot compositional learning. We tested our method on three public benchmarks, including both image and video datasets. Our results suggested that using external knowledge can help to better recognize novel interactions and even novel concepts of verbs and nouns. As a consequence, our model outperformed state-of-the-art methods on recognizing novel combination of seen concepts on all datasets. Moreover, our model demonstrated promising ability to recognize novel concepts. We believe that our model brings a new perspective to zero shot learning, and our exploration of using knowledge provides an important step for understanding human actions.

Acknowledgments. This work was supported by ONR MURI N000141612007, Sloan Fellowship, Okawa Fellowship to AG. The authors would like to thank Xiaolong Wang and Gunnar Sigurdsson for many helpful discussions.

References

1. Akata, Z., Perronnin, F., Harchaoui, Z., Schmid, C.: Label-embedding for attribute-based classification. In: CVPR (2013)
2. Andreas, J., Rohrbach, M., Darrell, T., Klein, D.: Learning to compose neural networks for question answering. In: NAACL (2016)
3. Andreas, J., Rohrbach, M., Darrell, T., Klein, D.: Neural module networks. In: CVPR (2016)
4. Biederman, I.: Recognition-by-components: a theory of human image understanding. Psychol. Rev. **94**(2), 115 (1987)
5. Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka Jr., E.R., Mitchell, T.M.: Toward an architecture for never-ending language learning. In: AAAI, pp. 1306–1313. AAAI Press (2010)
6. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: CVPR (2017)
7. Chao, Y.W., Wang, Z., He, Y., Wang, J., Deng, J.: HICO: a benchmark for recognizing human-object interactions in images. In: ICCV (2015)
8. Chen, X., Shrivastava, A., Gupta, A.: NEIL: extracting visual knowledge from web data. In: ICCV (2013)

9. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: CVPR (2005)
10. Delaitre, V., Fouhey, D.F., Laptev, I., Sivic, J., Gupta, A., Efros, A.A.: Scene semantics from long-term observation of people. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7577, pp. 284–298. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33783-3_21
11. Deng, J., et al.: Large-scale object classification using label relation graphs. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 48–64. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_4
12. Elhoseiny, M., Saleh, B., Elgammal, A.: Write a classifier: zero-shot learning using purely textual descriptions. In: ICCV (2013)
13. Fouhey, D., Wang, X., Gupta, A.: In defense of direct perception of affordances. In: arXiv (2015)
14. Fouhey, D.F., Delaitre, V., Gupta, A., Efros, A.A., Laptev, I., Sivic, J.: People watching: human actions as a cue for single-view geometry. Int. J. Comput. Vis. **110**(3), 259–274 (2014)
15. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Mikolov, T.: Devise: a deep visual-semantic embedding model. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) NIPS, pp. 2121–2129. Curran Associates, Inc. (2013)
16. Fu, Y., Hospedales, T.M., Xiang, T., Gong, S.: Transductive multi-view zero-shot learning. IEEE Trans. Pattern Anal. Mach. Intell. **37**(11), 2332–2345 (2015)
17. Gibson, J.: The Ecological Approach to Visual Perception. Houghton Mifflin, Boston (1979)
18. Guadarrama, S., et al.: YouTube2Text: recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In: ICCV (2013)
19. Gupta, A., Davis, L.S.: Objects in action: an approach for combining action understanding and object perception. In: CVPR (2007)
20. Gupta, A., Kembhavi, A., Davis, L.S.: Observing human-object interactions: using spatial and functional compatibility for recognition. IEEE Trans. Pattern Anal. Mach. Intell. **31**(10), 1775–1789 (2009)
21. Habibian, A., Mensink, T., Snoek, C.G.: Composite concept discovery for zero-shot video event detection. In: International Conference on Multimedia Retrieval (2014)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
23. Hoffman, D.D., Richards, W.A.: Parts of recognition. Cognition **18**(1–3), 65–96 (1984)
24. Jain, M., van Gemert, J.C., Mensink, T.E.J., Snoek, C.G.M.: Objects2Action: classifying and localizing actions without any video example. In: ICCV (2015)
25. Jain, M., van Gemert, J.C., Snoek, C.G.: What do 15,000 object categories tell us about classifying and localizing actions? In: CVPR (2015)
26. Jayaraman, D., Grauman, K.: Zero-shot recognition with unreliable attributes. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, pp. 3464–3472. Curran Associates, Inc. (2014)
27. Johnson, J., et al.: Inferring and executing programs for visual reasoning. In: ICCV (2017)
28. Kalogeiton, V., Weinzaepfel, P., Ferrari, V., Schmid, C.: Joint learning of object and action detectors. In: ICCV (2017)

29. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks (2017)
30. Krishna, R., et al.: Visual genome: connecting language and vision using crowdsourced dense image annotations. Int. J. Comput. Vis. **123**(1), 32–73 (2017)
31. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: CVPR (2009)
32. Leacock, C., Miller, G.A., Chodorow, M.: Using corpus statistics and wordnet relations for sense identification. Comput. Linguist. **24**(1), 147–165 (1998)
33. Li, X., Guo, Y., Schuurmans, D.: Semi-supervised zero-shot classification with label representation learning. In: CVPR (2015)
34. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
35. Liu, J., Kuipers, B., Savarese, S.: Recognizing human actions by attributes. In: CVPR (2011)
36. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 852–869. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_51
37. Mao, J., Wei, X., Yang, Y., Wang, J., Huang, Z., Yuille, A.L.: Learning like a child: fast novel visual concept learning from sentence descriptions of images. In: ICCV (2015)
38. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) NIPS, pp. 3111–3119. Curran Associates, Inc. (2013)
39. Miller, G.A.: WordNet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995)
40. Misra, I., Gupta, A., Hebert, M.: From red wine to red tomato: composition with context. In: CVPR (2017)
41. Norouzi, M., et al.: Zero-shot learning by convex combination of semantic embeddings (2014)
42. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: EMNLP (2014)
43. Rohrbach, M., Ebert, S., Schiele, B.: Transfer learning in a transductive setting. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) NIPS, pp. 46–54. Curran Associates, Inc. (2013)
44. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
45. Sadeghi, F., Kumar Divvala, S.K., Farhadi, A.: VisKE: visual knowledge extraction and question answering by visual verification of relation phrases. In: CVPR (2015)
46. Schlichtkrull, M., Kipf, T.N., Bloem, P., Berg, R.v.d., Titov, I., Welling, M.: Modeling relational data with graph convolutional networks. arXiv preprint arXiv:1703.06103 (2017)
47. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: ICCV (2017)
48. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: crowdsourcing data collection for activity understanding. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 842–856. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_31

49. Stark, L., Bowyer, K.: Achieving generalized object recognition through reasoning about association of function to structure. IEEE Trans. Pattern Anal. Mach. Intell. **13**, 1097–1104 (1991)
50. Thomason, J., Venugopalan, S., Guadarrama, S., Saenko, K., Mooney, R.: Integrating language and vision to generate natural language descriptions of videos in the wild. In: COLING (2014)
51. Wang, Q., Chen, K.: Alternative semantic representations for zero-shot human action recognition. In: Ceci, M., Hollmén, J., Todorovski, L., Vens, C., Džeroski, S. (eds.) ECML PKDD 2017. LNCS (LNAI), vol. 10534, pp. 87–102. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-71249-9_6
52. Wang, Q., Chen, K.: Zero-shot visual recognition via bidirectional latent embedding. Int. J. Comput. Vis. **124**(3), 356–383 (2017)
53. Xian, Y., Schiele, B., Akata, Z.: Zero-shot learning-the good, the bad and the ugly. In: CVPR (2017)
54. Xu, C., Hsieh, S.H., Xiong, C., Corso, J.J.: Can humans fly? Action understanding with multiple classes of actors. In: CVPR (2015)
55. Xu, X., Hospedales, T., Gong, S.: Semantic embedding space for zero-shot action recognition. In: ICIP (2015)
56. Xu, X., Hospedales, T.M., Gong, S.: Multi-task zero-shot action recognition with prioritised data augmentation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 343–359. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6
57. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: CVPR (2010)
58. Yu, X., Aloimonos, Y.: Attribute-based transfer learning for object categorization with zero/one training example. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6315, pp. 127–140. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15555-0_10
59. Zellers, R., Yatskar, M., Thomson, S., Choi, Y.: Neural motifs: scene graph parsing with global context. In: CVPR (2018)
60. Zhang, L., Xiang, T., Gong, S.: Learning a deep embedding model for zero-shot learning. In: CVPR (2017)