Learning Models for Actions and Person-Object Interactions with Transfer to Question Answering

Arun Mallya and Svetlana Lazebnik
University of Illinois at Urbana-Champaign

Abstract. In this paper, we propose a convolutional deep network model which utilizes local and global context through feature fusion to make human activity label predictions and achieve state-of-the-art performance on two different activity recognition datasets, the HICO and MPII Human Pose Dataset. We use Multiple Instance Learning to handle the lack of full person instance-label supervision and weighted loss to handle the unbalanced training data. Further, we show how expert knowledge from these specialized datasets can be transferred to improve accuracy on the Visual Question Answering (VQA) task, in the form of multiple choice fill-in-the-blank questions (Visual Madlibs). Specifically, we tackle two types of questions on person’s activity and person-object relationship and show improvements over generic features trained on the ImageNet classification task.

Keywords: Activity Prediction, Deep Networks, Context, VQA

1 Introduction

The task of Visual Question Answering or VQA has gathered a lot of interest over the past year with multiple datasets being released including the DAQUAR [1], VQA [2] and Visual Madlibs [3] among others. Various works have attempted to solve the problem with varying success by using an array of techniques such as memory networks, visual attention, recurrent networks and LSTMs [4,5,6,7]. All these methods use image features extracted from deep convolutional neural networks such as the VGG-16 network [8] trained on the ImageNet classification task [9], with or without finetuning during training for the VQA task. Questions in these VQA datasets are very open-ended and cover a wide array of concepts such as the presence or absence of various objects, animals, or humans in the image, counting, brand name recognition, human emotion, activity, and sport recognition, scene recognition and more. Given such a broad spectrum of questions, it is hard to believe that a generic one-size-fits-all high-level feature would emerge out of models trained on the ImageNet classification task.

Over the last few years, a number of valuable datasets and models targeting specific tasks such as scene recognition [10], age, gender, and emotion classification [11,12], human action recognition [13,14,15], etc. have been released. We
believe that in order to truly understand and answer questions about an image, we need to leverage and transfer the knowledge from these specialized datasets. Given a specific question, we would then be able to choose and blend features from multiple expert models or networks, thereby improving performance on VQA. Motivated by the recent availability of large activity recognition and VQA datasets centered around humans, in this paper, we show that transferring expert knowledge from a network trained on human activity prediction to the task of answering questions of the Madlibs dataset can not only improve performance, but also help interpret the why’s and how’s of the model’s decisions. Figure 1 shows the high-level outline of our method. We train deep networks on the HICO [14] and MPII [15] datasets to predict human activity. Our networks combine local and global information to make predictions using person boxes provided in MPII and detected boxes in HICO. We apply these networks on the questions of the Madlibs dataset to answer two different types of questions. Our contributions can be summed up as follows:

1. We propose simple deep convolutional models for predicting human activity labels by utilizing both local context from the person bounding box and global context from the whole image, and performing feature fusion for prediction as described in Section 3.
2. Using Multiple Instance Learning (MIL) to handle the lack of full person instance-label supervision and weighted loss to handle the unbalanced training data, we show that this simple model beats the previous state-of-the-art on two human activity recognition datasets as shown in Section 4.
3. By using image features extracted from our best performing networks on two different datasets with hundreds of category labels, we show that we can infuse expert knowledge into question answering models and improve accuracy on two types of questions from the Madlibs dataset targeting human activities and their interactions with objects, in Section 5.

2 Related Work

Recognition of human activities in still images, a popular task in computer vision, has wide applications in a world centered around humans and their daily activities. Various datasets have been released for single image based action recognition, including the older PASCAL VOC [16] and Stanford 40 Actions [17], and newer MPII Human Pose Dataset [15], COCO-A [18] and Humans Interacting with Common Objects (HICO) dataset [14]. The number of actions included in the newer datasets is an order of magnitude larger than the older ones, allowing us to learn vocabularies fit for general VQA. The HICO dataset is currently the largest human-object interaction dataset, consisting of nearly 50000 images belonging to 600 human-object interaction categories. Each category in the HICO dataset is composed of a verb-object pair, with objects belonging to the 80 object categories contained in the MS COCO dataset [19]. On the other hand, the MPII dataset comprises humans performing a 393 different activities including walking, running, skating, etc. in which they do not necessarily interact with objects.
Fig. 1: We train CNNs on the HICO and MPII datasets to predict human activity labels. Our networks use features from the full image and the person bounding boxes which are provided in the MPII dataset and detected in the HICO dataset. We then apply these networks on images of the Madlibs dataset, obtain activity labels, and use them to answer two types of multiple choice questions - about a person’s activity, and the relationship between a person and an object.

In this work, we train deep convolutional networks with simple architectures for activity recognition on these two datasets, and show that they outperform the previous state-of-the-art models on both datasets.

As activity labels for HICO are provided for the image as a whole, aggregated over all people in the image instead of being associated with specific people in the image, we formulate a way to disambiguate activity label assignment over the people in the image by using the Multiple Instance Learning (MIL) formulation [20]. The MIL framework provides a way to learn models when fully supervised data is not available at train time. Instead of receiving a set of individually labeled ‘instances’, a set of labeled ‘bags’ is provided to the learner. Each bag is labeled negative if all the instances inside it are labeled negative, and labeled positive if it contains at least one instance which has a positive label. The MIL framework has been widely used in computer vision in problems where training data is often weakly supervised or incompletely labeled such as object detection [21, 22], semantic segmentation [23, 24], etc. In this work, we adapt the MIL framework to our problem by treating each person in the image as an ‘instance’ and the image, which contains one or more people in it, as a ‘bag’. The exact formulation of our learning procedure is explained in Section 3.2.

Contextual information outside the tight box around an object has shown to be useful in tasks such as object detection & classification [25, 26, 27] and segmentation [28]. The recent R*CNN network of Gkioxari et al. [29] used contextual cues by choosing a second box that overlapped the bounding box of the person to some defined extent and provided the strongest evidence of a particular action.
being performed to obtain state-of-the-art results on multiple action recognition datasets. The choice of the supplementary box was guided by a MIL formulation of the deep network. A simpler version which uses the entire image instead of a chosen box, the Scene-RCNN, was also proposed. Exploiting contextual information outside a proposed bounding box by using spatial recurrent neural networks also helped the *Inside-Outside Net* (ION) of Bell et al. [30] achieve very good object detection performance on the PASCAL VOC and MS COCO datasets. In our work, while making predictions for a single person, we employ a simple framework that combines features from both the entire image and the bounding box of the person under consideration, as detailed in Section 3.1.

We build on the *Fast RCNN* (FRCN) [31] based on the VGG-16 architecture [8]. The major contribution of the FRCN was a new adaptive max pooling layer, referred to as the ROI pooling layer, that replaces the standard max pooling layer (pool5) after the set of the first five convolutional layers. This layer takes in a list of bounding boxes, referred to as Regions Of Interest (ROI) and outputs a feature map of fixed spatial size (commonly $7 \times 7$) for each input ROI. These pooled features are then fed to the fully connected layers. The Fast RCNN is very efficient as it requires the computationally intensive convolutions to be run just once on the entire image following which ROI pooling features can be extracted for each region of interest in the image. This architecture allows for backpropagation through the ROI pooling layer unlike the SPPNet [32].

### 3 Method

#### 3.1 Network Architecture

When we focus on predicting the actions of a single person, it is beneficial to use cues from both the bounding box of the person and other parts of the image outside the bounding box [26,29]. By looking at just a tight box around the person of interest, we might severely crop out or totally miss the object that the person is interacting with or ignore strong cues present in the rest of the image. Previous work suggests the use of latent context boxes [29], multiresolution or zoom-out features [30,33] and complex 2-D recurrent structures [30]. We explored using latent boxes but found their performance to be lacking possibly due to the large output space and sampling during training. Similarly, we could not obtain good results with multiresolution features owing to overfitting. The simple method of using context from the entire image surprisingly outperformed these powerful methods.

While making predictions for a specific person, we extract features from both the bounding box of the person and the entire image. During the forward pass of the network, we use two ROIs per person in the image: the bounding box of the person, and the full image. We also experimented with using an expanded person bounding box instead of the full image, but found the full image to always work better. The ROI Pooling layer produces a feature of 512 channels and spatial size $7 \times 7$ for each ROI. In order to achieve good performance on smaller datasets, previous work [34,35,30] has emphasized the importance of re-using weights from
(a) Fusion-1 network: Features from bounding box and full image are stacked and a $1 \times 1$ convolution is used for dimensionality reduction.

(b) Fusion-2 network: Features from bounding box and full image undergo dimensionality reduction using a $1 \times 1$ convolution each, and are then stacked.

Fig. 2: The networks used in our paper extracts Region Of Interest (ROI) features \[31\] of dimension $512 \times 7 \times 7$ from both the person bounding box and the full image. We try two variants as shown above which differ in the order in which feature dimensionality reduction and feature concatenation are performed. The resulting feature is fed into the $fc6$ and following layers of the VGG-16 network.

We explore two ways of combining the two ROI features, one from the person bounding box, and the other from the full image, through stacking and dimensionality reduction. In the first, referred to as Fusion-1, we stack features from the bounding box and the entire image along the channel dimension and obtain a feature of size $1024 \times 7 \times 7$. A convolutional layer of filter size $1 \times 1$ is used to perform dimensionality reduction of channels from 1024 to 512, while keeping the spatial size the same. In the second, referred to as Fusion-2, we first perform dimensionality reduction on the two ROI features individually to reduce the number of channels from 512 to 256 each, and then stack the two dimensionality reduced features to obtain an input of size $512 \times 7 \times 7$ for the $fc6$ layer. Figure 2 depicts the two feature fusion network architectures used in our work.

This architecture differs from the $R^* CNN$ in two major ways: we do not explicitly try to find a box or set of boxes which provides support for a particular label, and we combine features before prediction instead of independently performing prediction using the two features and then combining predictions. The results presented in Section 4 also show that independently performing predictions using the bounding box and the full image, and then averaging predictions (like the Scene-RCNN from \[29\]) performs worse than combining features before
prediction. Further, our architecture is faster to run due to the lack of the need to sample boxes during training and testing.

### 3.2 Multiple Instance Learning for Label Prediction

In the HICO dataset, if at least one of the people in the image is performing an action, the label is marked as positive for the image. As our architecture makes predictions with respect to a person bounding box (Section 3.1), we treat the assignment of labels to different people as latent variables and try to infer the assignment during end-to-end training of the network. For an image \( I \), let \( D \) be the set of all person bounding boxes in the image. Using our network described above which takes as input an image \( I \) and a person bounding box \( d \in D \), we obtain the score of an action \( \alpha \) for the image as follows:

\[
score(\alpha; I) = \max_{d \in D} score(\alpha; d, I)
\]

where \( score(\alpha; d, I) \) is the score of action \( \alpha \) for the person \( d \) in image \( I \). The predicted label for the action can be obtained by passing the score through a logistic sigmoid or softmax unit as required. The max operator enforces the constraint that if a particular action label is active for a given image, then at least one person in the image is performing that action, and when a particular action label is inactive for a given image, then no person in the image is performing the action. During the forward pass, the score and thus the label for the image are predicted using the above relationship. The predicted label is compared to the groundtruth label in order to compute the loss and gradients for backpropagation.

### 3.3 Weighted Loss Function

Mostajabi et al. [33] showed that use of an asymmetric weighted loss helps greatly in the case of an unbalanced dataset. For the HICO dataset, we have to learn 600 independent classifiers per image and this makes for a highly unbalanced scenario, with the number of negative examples greatly outnumbering the positive examples, even for the most populous categories (an average negative to positive ratio of 6000:1, worst case of 38116:1). We thus compute a weighted cross-entropy loss in which positive examples are weighted by a factor of \( w_p \) and negative examples by a factor of \( w_n \), which may vary from class to class. Given a training sample \((I, D, y)\) consisting of an image \( I \), set of person bounding boxes or detections \( D \), and ground truth action label vector \( y \in \{0, 1\}^C \) for \( C \) independent classes, the network produces probabilities of actions being present in the image by passing predictions through a sigmoid activation unit. For any given training sample, the training loss on network prediction \( \hat{y} \) is thus given by

\[
loss(I, D, y) = \sum_{i=1}^{C} w_p^i \cdot y^i \cdot \log(\hat{y}^i) + w_n^i \cdot (1 - y^i) \cdot \log(1 - \hat{y}^i)
\]

In our experiments, we set \( w_p = 10 \) and \( w_n = 1 \) for all classes for simplicity.
4 Activity Prediction Results

Datasets. We train and test our system on two different activity classification datasets: HICO [14] and the MPII Human Pose Dataset [15]. The HICO dataset contains labels for 600 human-object interaction activities, any number of which might be simultaneously active for a given image. Labels are provided at the image level even though each image might contain multiple person instances, each performing the same or different activities. The labels can thus be thought of as an aggregate over labels of each person instance in the image. Thus, when any model uses person bounding boxes on the HICO dataset, we also have to use the Multiple Instance Learning (MIL) formulation as we only have image level labels. As the person bounding boxes are not provided with the HICO dataset, we run the Faster-RCNN detector [36] with the default confidence threshold of 0.8 on all the train and test images. The obtained person bounding boxes are thus not perfect and might have wrong or missing annotations. The HICO training set contains 38,116 images and the test set contains 9,658 images. The HICO dataset is highly unbalanced with 51 out of 600 categories having just 1 positive example in the entire train set.

The MPII dataset contains labels for 393 actions, but unlike HICO, each image only has a single label. All marked person instances inside an image are assumed to be performing the same task. Unlike the HICO dataset, in the MPII dataset each person bounding box is known to exhibit the activity assigned to the image and we do not necessarily need to use the MIL framework and can take advantage of the extra training data available by training on each person instance separately. Further, unlike HICO, the groundtruth bounding box is available for each instance in the training set, but only a single point inside the bounding box is provided for each instance in the test set. We thus run the Faster-RCNN detector to detect people in the test set. The training set consists of 15,200 images and 22,900 person instances and the test set has 5,709 images. Similar to HICO, the training set is unbalanced and the number of positive examples for a label ranges from anywhere between 3 to 476 instances.

Models and Baselines. On the HICO dataset, we compare our network described in the previous section with the VGG-16 network trained on just the person bounding boxes and just the full image as well as train the state-of-the-art R*CNN. For all of our networks, except the R*CNN, we use a learning rate of $10^{-5}$, decayed by a factor of 0.1 every 30000 iterations. For the R*CNN, we use the recommended setting from [29] of a learning rate of $10^{-4}$, with a lower and upper intersection over union (IoU) bound for secondary regions of 0.2 and 0.75 and sample 10 secondary regions per person bounding box during a single training pass. We train all networks for 60000 iterations with a momentum of 0.9. Further, all networks are finetuned till the conv3 layer as in previous work [29,31]. We use a batch size of 10 images, resize images to a maximum size of 640 pixels, and sample a maximum of 6 person bounding boxes per image in order to fit the network in the GPU memory during training with MIL.

On the MPII dataset, we compare our network with previously published baselines from Pischulin et al. [15] and Gkioxari et al. [29]. Our networks are
trained with a learning rate of $10^{-4}$ with a decay of 0.1 every 12000 iterations, for 40000 iterations. We only finetune till the $fc6$ layer on the MPII dataset due to the lesser amount of data available. Using the weighted loss was not found to make a difference on the softmax loss of MPII dataset and the reported results did not use weighted loss.

| Method                     | Full Im. | Bbox | MIL | Wtd. Loss | mAP |
|----------------------------|----------|------|-----|-----------|-----|
| a) AlexNet+SVM [14]        | ✓        |      |     |           | 19.4|
| VGG-16                     | ✓        |      | ✓   |           | 29.4|
| VGG-16                     | ✓        | ✓    | ✓   | ✓         | 14.6|
| VGG-16, R*CNN              | ✓        | ✓    | ✓   | ✓         | 28.5|
| VGG-16, Scene-RCNN         | ✓        | ✓    | ✓   | ✓         | 29.0|
| b) VGG-16, fusion-1        | ✓        | ✓    | ✓   | ✓         | 33.6|
| VGG-16, fusion-1           | ✓        | ✓    | ✓   | ✓         | 36.0|
| VGG-16, fusion-2           | ✓        | ✓    | ✓   | ✓         | 33.8|
| VGG-16, fusion-2           | ✓        | ✓    | ✓   | ✓         | 36.1|

Table 1: Performance of various networks on the HICO person-activity dataset. Our networks which fuse features from both person bounding box and the full image, outperform other architectures. Note that usage of the Bounding Box (Bbox) necessitates the usage of Multiple Instance Learning (MIL).

Activity Prediction Performance. As the HICO dataset is new, the only published baseline [14] uses the AlexNet as shown in Table 1a. Using the VGG-16 network improves upon AlexNet by 10 mAP (Table 1b). The VGG-16 network that uses just the person bounding box to make predictions with MIL, as described in Section 3.2 performs poorly with just 14.6 mAP. This is expected behavior as the object that the person is interacting with is often not completely inside the tight bounding box of the person. Interestingly, the R*CNN network [29] which holds the state-of-the-art performance on the PASCAL VOC action classification task and uses latent context boxes performs slightly worse than the full image VGG network. There might be a number of reasons why this behavior is exhibited. The R*CNN has to use Multiple Instance Learning twice during learning - once for finding the secondary box that provides the most evidence for a particular action label, and then again while aggregating over the multiple person instances in the image. This network learns by sampling 10 boxes per person instance during each pass of training and finding the right box for each of the 600 actions might be difficult. The Scene-RCNN which uses the entire image as the selected secondary box needs MIL just once, over the multiple person instances in the image, and performs marginally better than R*CNN. In both the R*CNN and Scene-RCNN, action score predictions are made independently using the person bounding box and the secondary box and then summed. Fine-grained classification of actions from a secondary box, which may or may not have a high overlap with the interacting person, would be hard.

Our architecture, though might seem similar to Scene-RCNN, is actually very different. Unlike the Scene-RCNN, our network fuses features from both the per-
son bounding box and the entire image before making a prediction. The Scene-RCNN has to perform the task of predicting activity score from just the tight bounding box. Using our fusion networks, we immediately see improvements over the full image network (Table 1c). Using a weighted loss that penalizes mistakes on positive examples more heavily as described in Section 3.2 helps push the mAP higher by about 2.5 mAP for both our networks. The Fusion-2 network which performs dimensionality reduction before feature local and global feature concatenation has a slight edge probably due to lower number of parameters (Fusion-1 has $1024 \times 512$ parameters for dimensionality reduction and Fusion-2 has $2 \times 512 \times 256$, lesser by a factor of 2) and separated dimensionality reduction.

| Method | mAP |
|--------|-----|
| Dense Trajectory + Pose | 5.5 |
| VGG-16, R*CNN | 26.7 |
| VGG-16, fusion-1, MIL over person instances | 31.68 |
| VGG-16, fusion-2, MIL over person instances | 31.89 |
| VGG-16, fusion-1 | 32.06 |
| VGG-16, fusion-2 | 32.24 |

Table 2: Results on the MPII test set. Feature fusion fares better than previous work, and full supervision during training gives gains over weakly supervised Multiple Instance Learning (MIL).

Table 2 compares our networks with the previous results on the MPII dataset. We observe a trend similar to HICO, wherein our feature fusion networks outperform previous methods, with Fusion-2 having a lead over the Fusion-1 network. As the MPII dataset has multiple person instances in an image to which the activity label applies, this gives us a chance to examine the effect of our MIL framework. If we assume that the assignment of labels to the people in the image is unknown, and use the MIL framework, we see a small dip in performance as opposed to using the information that the label applies to each person in the image (last two rows of Table 2). The latter gives us more training data along with full supervision during training and improves over MIL by around 0.4 mAP.

Qualitative Results. In Figure 3, we display some of the positive predictions of our best performing network trained using MIL on the HICO dataset. In spite of the lack of explicit supervision of which labels map onto a specific person instance, by going over a large number of examples, the network learns to reasonably assign labels to the correct person instance. It is also interesting to note a few minor mistakes made by the network in the above examples: in the top left image, the network confuses the tower in the background for a clock tower, and assigns the label ‘no_interaction-clock’ label to one of the people. In the middle image of the second row, there is a false person instance (marked in red) due to the reflection in the glass found by the person detector we use.

Figure 4 shows some of the failures of our system. Unusual use-cases of an object such as swinging around a backpack can confuse the deep network into misclassifying the object as in the leftmost image. Since our system relies on detected people as the groundtruth information is not available, we find cases
Fig. 3: Predictions on the HICO test set where the Multiple Instance Learning (MIL) using weak label supervision has learnt to assign labels to the right person instance. Note that this includes person instances which are far away, close by, as well as have significant overlap with each other.

Fig. 4: Examples where our system fails. Incorrect classification of objects/actions, wrong interacting person detection, and inability to assign labels to correct person instances due to weak supervision and sampling are common issues with our network trained using the MIL framework.

where we either miss or produce false positives for people instances or label the wrong instances as shown in the middle image. Lastly, one drawback of the weakly supervised MIL framework is that it is unable to distinguish labels in a crowded scenario, especially when the crowd occurs only in specific settings. As in the case of the rightmost image, such crowds are normally only found in baseball games (probably a dataset bias) and the framework is unable to tell to whom the labels should be assigned.

5 Visual Question Answering Results

Dataset and Tasks. In order to show that domain information is essential and useful for good performance in Visual Question Answering (VQA) problems
instead of a general one-size-fits-all feature, we evaluate the performance of features extracted by our networks on questions from the Madlibs dataset [3]. We focus on two types of questions that specifically target people and objects, their activities and interactions. The first type, ‘Person’s Activity’ asks us to choose an option which best describes the activity of the indicated person/people, while the second type, ‘Pair’s Relationship’ asks us to describe the relationship between the indicated person and object(s). All the examples contain a list of one or more people that the question is about and 4 possible answer choices, of which one is the correct answer (or most correct answer, in case of confusing options). The prompt is fixed for all questions of a particular type - ‘The person/people is/are ___’ and ‘The person/people is/are ___ the object(s)’ for the person’s activity and pair’s relationship tasks respectively. The dataset contains 26528 and 30640 training examples, and 6501 and 7595 testing examples for the person’s activity and pair’s relationship tasks respectively. Depending on the difficulty of the distractor options provided, questions are divided into two categories - Easy and Hard. Options in the hard category are often quite confusing, with even humans disagreeing on the correct answer. Thus, the performance on filtered hard questions, a subset of the hard questions on which human annotators agree with the ‘correct’ answer at least 50% of the times, is also measured. We created our own held out validation set consisting of 10% of the train images by following the distractor choice generating procedure used in the original paper of [3].

Models and Baselines. We use the normalized Canonical Correlation Analysis (nCCA) [37] model which has been shown to be significantly better than the original CCA models and learn a joint embedding space of dimensionality 300 to which the image and the choice features are to be mapped. Given a question, we select the choice that has the highest cosine similarity with the image features in the joint embedding space as the predicted answer.

In all our experiments, we represent each of the choices by the average of the word2vec features [38] of the words in the choice. The word2vec features for any word, and thus the choice, are of length 300. We use the publicly available word2vec system [39] for extracting word features. In the case that a word was out of the vocabulary of the trained system, we use an all-zero feature to represent the word. We evaluate and compare performance obtained by using three different types of images features. The first type of feature is obtained by passing the entire image, resized to $224 \times 224$ pixels, through the vanilla VGG-16 network and extracting the fc7 activations, and this serves as the baseline, similar to the original work of Yu et al. [3]. The second type of feature is obtained by passing the entire image through the activity prediction network that uses full image inputs and is trained on the HICO dataset. We try using features extracted from both the fc7 layer of length 4096 and the class label predictions of length 600. Lastly, we extract features by using the architecture that uses both the bounding box and the whole image to make predictions (as detailed in Section [3]). As the question types we tackle from the Madlibs dataset target one or more specific people in the image, we feed in the person bounding boxes as ROIs to our network to obtain class label predictions. In the case that a partic-
ular question targets more than one people, we perform max pooling over the class label predictions of the distinct people to obtain a single feature vector. For the class label features, we found it necessary to use the predictions before passing them through the logistic sigmoid/softmax as the squashing saturated the activations too close to 0 or 1. Using the validation set, we determined that a regularization parameter value of 0.01 and 0.001 for the CCA model worked best with the $fc7$ and class score features respectively.

| Dataset:Network                  | Feature | Person’s Activity | Pair’s Relationship |
|---------------------------------|---------|-------------------|--------------------|
|                                 |         | Easy   | Hard  | Fil. H. | Easy   | Hard  | Fil. H. |
| ImageNet:VGG-19 [3]             | fc7     | 80.7   | 65.4  | 68.8    | 63.0   | 54.3  | 57.6    |
| ImageNet:VGG-16                 | fc7     | 80.79  | 65.14 | 67.73   | 71.45  | 51.47 | 56.28   |
| HICO:VGG-16, Full Im.           | cls_score| 86.03  | 68.74 | 72.06   | 77.25  | 54.10 | 59.77   |
| HICO:VGG-16, Full Im.           | fc7     | 86.54  | 69.14 | 72.39   | 77.96  | 55.76 | 61.03   |
| HICO:Fusion-2                   | cls_score| 86.66  | 70.05 | 73.46   | 78.29  | 55.52 | 61.39   |
| MPII:Fusion-2                   | cls_score| 83.23  | 68.11 | 70.89   | 72.81  | 52.75 | 57.68   |
| HICO+MPII:Fusion-2              | cls_score| **87.57**| **71.13**| **74.45**| **78.50**| **56.17**| **62.06**|

Table 3: Performance on Madlibs. Using image features learned for activity prediction helps improve accuracy, with features from the network that uses bounding box and full image giving the best performance. (Fil. H. ≡ Filtered Hard)

**Question Answering Performance.** The first row of Table 3 contains the accuracy numbers obtained from the paper of Yu et al. [3]. Note that there are some differences in the accuracies, probably owing to the different features used (VGG-16 v/s VGG-19), toolboxes used to learn the CCA models (we used the openly available toolbox of [40]), and effect of hyperparameters. We consider our implementation as the baseline, so as to compare methods based on VGG-16 and trained using the same toolbox. As presented in the second row of Table 3, using the vanilla VGG features gives an accuracy of 80.79% and 71.45% on the Easy Person Activity and Easy Pair Relationship questions respectively. By extracting full image features from our network trained on human-object interaction prediction on the HICO dataset, we immediately obtain gains of around 6 – 7% on the Easy task for both question types (rows 2-3). For the person’s activities, it is interesting to note that using the class label predictions of much lower dimensionality of 600 gives performance comparable (within 0.5%) to the higher dimensional $fc7$ features. Using targeted person bounding boxes to make HICO class label predictions helps improve the performance further. Lastly, we obtained the best performance by using class label predictions from both the HICO and the MPII datasets. This performance improvement can be explained by the fact that MPII categories are complementary to those of HICO, especially in the case that a person is not interacting with an object. Compared to the baseline, we obtain an improvement of 6.8% on the easy person’s activity task, and 7.05% on the easy pair’s relationship task. For the hard person’s activity task, we obtain an improvement of 6% and 6.7% on the unfiltered and filtered questions.
respectively. For the hard pair’s relationship task, we obtain an improvement of 4.7% and 5.8% on the unfiltered and filtered questions respectively.

Fig. 5: Correctly answered questions of the person activity type. The one or more subjects of the question are highlighted in each image. The left column below each image shows the answer choices for each question, with the correct answer marked in red. The right column shows the activity labels predicted by our best network. The predicted labels often align very closely to the correct choice.

Qualitative Results. In Figure 5, we show a range of correctly answered multiple choice questions by the best performing input features. Under each image, the left column contains the four available choices, with the correct answer highlighted in red. The right column shows the top labels from the HICO dataset predicted for the one or more people targeted by the question. By examining these high-level labels, we can gain intuition into the choices of our model as these are easily interpretable unlike \( fc7 \) features of the VGG-16 network.

In the top left image of Figure 5, the question targets multiple people and the labels aggregated over the people using max pooling correctly predicts the activity of sitting at and eating at the dining table. In the middle image of the first row, targeting the skateboarder gives a high score for skateboard-related activities, and a much lower score for the bicyclist in the background. It is interesting to note that in the rightmost image in the first row, the network also correctly predicts the labels for ‘wear, carry-backpack’, along with ‘ride, straddle-horse’. Further, the middle and right image in the bottom row show that our predictions change depending on the target bounding box - the ‘hold-book’ label has a much higher probability for the boy on the right, even though the network
1.00, no_interaction-clock!
0.73, no_interaction-vase!
0.38, wear-tie!
0.12, sit_on-chair!
holding a picture!
looking
standing
1.00, no_interaction-clock!
0.35, wear-tie!
0.12, sit_on-chair!
shopping
sitting on the toilet!
taking a selfie!
taking a photo!
0.33, type_on-keyboard!
0.27, no_interaction-oven!
0.15, load-truck!
0.15, hold-keyboard!
parasurfing
walking!
holding board!
talking!
0.84, hold-surfboard!
0.81, hold-kite!
0.77, carry-surfboard!
0.22, fly-kite!

Fig. 6: Incorrectly answered questions of the person activity type. The correct choice is marked in red, and the predicted answer in blue. Failure modes mainly belong to three classes as illustrated (left to right): correct predictions but unfamiliar object (‘picture’), incorrect predictions (‘sink’ missed), and a mix of the first two - partly correct predictions and unfamiliar setting.

was trained using weak supervision and MIL, as detailed in Section 3. Note that a network which uses a full image input would produce a single feature vector for the whole image, and not separately for each person instance.

Figure 6 displays some of the common failure modes of our system. As in the leftmost image, even though the system makes correct predictions, the specific object in the image (‘picture’) was absent in the HICO and MPII datasets and the labels offer absolutely no useful information about the object. This results in the nCCA model ranking an unrelated choice as the best. The network can also make a wrong prediction in hard examples such as the middle image due to confusing visual cues. In the rightmost image, the choices are rather hard and confusing as the person is indeed holding onto a kite as well as a surfboard in an activity best described as ‘parasurfing’ or ‘windsurfing’.

6 Conclusion

In this paper, we developed effective models which exploit local and global context to make predictions and showed how Multiple Instance Learning could be used to train models in cases where only weak supervision was available. Even though we used simple contextual representation from the whole image, we obtained state-of-the-art performance on two different datasets, outperforming sophisticated models like the R∗CNN. In future work, we hope to explore contextual models with newer techniques such as attention and recurrent networks.

We have also shown how transfer learning of expert knowledge can be used to improve the performance on VQA tasks. While we demonstrated this on person activities, we envision a system which would have access to many more input features such as person attributes, detected objects, scene information, etc. and would combine them based on the question and image provided. Examining the routing of such information would also help us better understand and analyze the performance of VQA systems, in the way that high-level class activity labels helped us gain insight into the behavior and mistakes of our models.
References

1. Malinowski, M., Fritz, M.: A multi-world approach to question answering about real-world scenes based on uncertain input. In: NIPS. (2014)
2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: ICCV. (2015)
3. Yu, L., Park, E., Berg, A.C., Berg, T.L.: Visual madlibs: Fill in the blank image generation and question answering. In: ICCV. (2015)
4. Xu, H., Saenko, K.: Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. arXiv preprint arXiv:1511.05234 (2015)
5. Zhou, B., Tian, Y., Sukhbaatar, S., Szlam, A., Fergus, R.: Simple baseline for visual question answering. arXiv preprint arXiv:1512.02167 (2015)
6. Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., Xu, W.: Are you talking to a machine? dataset and methods for multilingual image question answering. arXiv preprint arXiv:1505.05612 (2015)
7. Shih, K.J., Singh, S., Hoiem, D.: Where to look: Focus regions for visual question answering. (2016)
8. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556 (2014)
9. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. IJCV (2015)
10. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: NIPS. (2014)
11. Levi, G., Hassner, T.: Age and gender classification using convolutional neural networks. In: CVPR Workshop. (2015)
12. Levi, G., Hassner, T.: Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In: ACM ICML. (2015)
13. Majji, S., Bourdev, L., Malik, J.: Action recognition from a distributed representation of pose and appearance. In: CVPR. (2011)
14. Chao, Y.W., Wang, Z., He, Y., Wang, J., Deng, J.: Hico: A benchmark for recognizing human-object interactions in images. In: ICCV. (2015)
15. Pishchulin, L., Andriluka, M., Schiele, B.: Fine-grained activity recognition with holistic and pose based features. In: GCPR. (2014)
16. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV (2010)
17. Yao, B., Jiang, X., Khosla, A., Lin, A.L., Guibas, L., Fei-Fei, L.: Human action recognition by learning bases of action attributes and parts. In: ICCV. (2011)
18. Ronchi, M.R., Perona, P.: Describing common human visual actions in images. In: BMVC. (2015)
19. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. (2014)
20. Maron, O., Lozano-Pérez, T.: A framework for multiple-instance learning. In: NIPS. (1998)
21. Zhang, C., Platt, J.C., Viola, P.A.: Multiple instance boosting for object detection. In: NIPS. (2005)
22. Hoffman, J., Pathak, D., Darrell, T., Saenko, K.: Detector discovery in the wild: Joint multiple instance and representation learning. In: CVPR. (2015)
23. Vezhnevets, A., Buhmann, J.M.: Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning. In: CVPR. (2010)
24. Pinheiro, P.O., Collobert, R.: From image-level to pixel-level labeling with convolutional networks. In: CVPR. (2015)
25. Torralba, A., Murphy, K.P., Freeman, W.T.: Using the forest to see the trees: exploiting context for visual object detection and localization. Communications of the ACM (2010)
26. Li, C., Parikh, D., Chen, T.: Extracting adaptive contextual cues from unlabeled regions. In: ICCV. (2011)
27. Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A.: The role of context for object detection and semantic segmentation in the wild. In: CVPR. (2014)
28. Shotton, J., Winn, J., Rother, C., Criminisi, A.: Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. IJCV (2009)
29. Gkioxari, G., Girshick, R., Malik, J.: Contextual action recognition with rcnn. In: ICCV. (2015)
30. Bell, S., Zitnick, C.L., Bala, K., Girshick, R.B.: Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. CoRR abs/1512.04143 (2015)
31. Girshick, R.: Fast r-cnn. In: ICCV. (2015)
32. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. TPAMI (2015)
33. Mostajabi, M., Yadollahpour, P., Shakhnarovich, G.: Feedforward semantic segmentation with zoom-out features. In: CVPR. (2015)
34. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR. (2014)
35. Agrawal, P., Girshick, R., Malik, J.: Analyzing the performance of multilayer neural networks for object recognition. In: ECCV. (2014)
36. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NIPS. (2015)
37. Gong, Y., Ke, Q., Isard, M., Lazebnik, S.: A multi-view embedding space for modeling internet images, tags, and their semantics. IJCV (2014)
38. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS. (2013)
39. Google: Word2vec trained model. [https://code.google.com/archive/p/word2vec/] Accessed: 8 March 2016.
40. Klein, B., Lev, G., Sadeh, G., Wolf, L.: Associating neural word embeddings with deep image representations using fisher vectors. In: CVPR. (2015)