

Supplementary Material for Motion Guided Attention Fusion to Recognize Interactions from Videos

– Supplementary Material for ICCV Submission ID 7911 –

1. Introduction

In this supplementary material, we provide details necessary for reproducing our experimental results on the Something-Else dataset in Section 2. Then, we report more details on the experimental set up of our IKEA-Assembly experiments in Section 3. As part of the supplementary material submission package, we also included the implementation of our approach in PyTorch which we will publicly release upon publication of this work.

2. Additional Details on the Something-Else Experiments

2.1. Model & optimization settings

All our experiments use the same settings for consistency. Given GPU availability, we used the largest possible batch size and adjusted the learning rate according to [3] when the batch size changed.

We use the SlowFast [2] model pretrained on Kinetics-400. We use the ResNet-50 backbone for the SlowFast model with $\beta = 1/8$ for the channel ratio and $\alpha = 8$ for the speed ratio. We use input jitter in range [224, 256] during training.

We optimize using SGD with momentum of 0.9 and weight decay of $1e-4$. We use a batch size of 16 and a learning rate of 0.01. We train for 100 epochs with a cosine learning rate decay schedule. We use a learning rate warm up schedule for 34 epochs starting from an initial learning rate of 0.001. We do not perform evaluations under the ‘accurate’ setting for fair comparison to other methods. This means we only use a single test crop (instead of 10 crop testing) and sample 8 frames (instead of 16) per video. In practice, using the ‘accurate’ evaluation setting leads to few extra points in classification performance but we omit it as other state of the art baselines do not use this setting. All our models are trained with the PyTorch framework on varying GPU hardware: most of the results are trained and generated using two TITAN-V GPUs.

2.2. Object detections

We provide more detail on the object detections used in this work. For all action models that use object detections, we use identical detections as [6] such that any performance difference is not caused by difference in quality of object detections. We also use the same object tracks released by [6] in all our experiments for the Something-Else dataset.

We provide more detail on our per-frame object detection representation discussed in Section 3.2. For Something-Else experiments, we assume at most 5 objects per frame such that $D = 5$. As described in Section 3.2, we represent each object as a vector of four numbers corresponding to its bounding box coordinates. We represent each frame as a concatenation of at most five ($D = 5$) object positions leading to a $4 \times 5 = 20$ dimensional vector. For example, a video of length 10 would be a 2D tensor of shape 20×10 . If there are fewer than five objects, we zero-pad the corresponding dimensions. If there are more than five objects, we use the detection score to pick the top five. In terms of ordering of objects, hand/person is always the first object. The ordering of remaining objects is arbitrary and is determined by object detections provided by [6].

3. Additional Details on the IKEA-Assembly Experiments

Regarding the experiments for the IKEA-Assembly dataset [1], all model and optimization settings are identical to those used in the Something-Else experiments. We will first provide details regarding the object detections used for this dataset. Then, we describe in more detail the original and the compositional tasks defined for the IKEA-Assembly dataset.

3.1. Object detections

The IKEA-Assembly dataset released manual instance segmentation annotation for 1% of the frames. The manually annotated frames are used to then train a Mask RCNN model [4] with ResNet-50 [5] backbone with Feature Pyramid networks structure. The object segmentation networks are trained using the Detectron2 framework¹ and the predictions from the trained models are released by the authors of [1].

We use the bounding box predictions of the trained Mask RCNN model released by the authors of [1] as is. This means all our results reported in Section 5 of the main manuscript are results using object predictions instead of ground truth object locations. This goes to show that our approach works well even when ground truth locations of objects are unknown by leveraging object location predictions.

3.2. The original and compositional tasks

The authors of the IKEA-Assembly dataset defined an action classification set up where the train/test split was defined by the data capture environment. The sets of environments that appear in train and test sets are disjoint but there is an overlap in the action categories themselves. For example, the model gets to observe an instance of `flip table` during training and is tested on a different `flip table` instance captured from a different environment. Table 1a lists all the actions that are defined for the action classification task which leads to a 33-way classification set up.

We provide more details regarding the compositional task defined for the IKEA-Assembly dataset that we discuss in 5.2 of the main paper. The main limitation of the original task for this dataset is that it does not explicitly test for the ability of the model to recognize an action performed with a previously unseen object. We introduce the compositional task to test for the model’s ability to generalize to unseen verb-noun compositions. For example, in this set up, the model gets to observe `flip table` instances at training time but is tested on an instance of `flip shelf`. The task is then to predict the verb (ie. `flip`) of the given sample. Table 1b describes the compositional train/test split of actions that we defined. The compositional split leads to a 6-way classification of predicting the verbs of an action.

References

1. Y. Ben-Shabat, X. Yu, F. Saleh, D. Campbell, C. Rodriguez-Opazo, H. Li, and S. Gould. The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020. 2
2. C. Feichtenhofer, H. Fan, J. Malik, and K. He. Slowfast networks for video recognition. *International Conference on Computer Vision*, pages 6202–6211, 2018. 1
3. P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arxiv*, 06 2017. 1
4. K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 2
5. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2
6. J. Materzynska, T. Xiao, R. Herzig, H. Xu, X. Wang, and T. Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

¹ <https://github.com/facebookresearch/Detectron>

ID	Action Label
0	NA
1	Align leg
2	Align side panel
3	Attach back panel
4	Attach side panel
5	Attach shelf
6	Flip shelf
7	Flip table
8	Flip table top
9	Insert pin
10	lay down back panel
11	lay down bottom panel
12	lay down front panel
13	lay down leg
14	lay down shelf
15	lay down side panel
16	lay down table top
17	Other
18	Pick up back panel
19	Pick up bottom panel
20	Pick up front panel
21	Pick up leg
22	Pick up pin
23	Pick up shelf
24	Pick up side panel
25	Pick up table top
26	Position drawer
27	Push table
28	Push table top
29	Rotate table
30	Slide bottom panel
31	Spin leg
32	Tighten leg

(a) Action labels defined for the original action classification task from the IKEA-Assembly dataset.

Train set		Test set	
Verb ID	Action label	Verb ID	Action label
0	Align leg	0	Align side panel
1	Attach back panel	1	Attach side panel
1	Attach shelf		
2	Flip shelf	2	Flip table
2	Flip table top		
3	Lay down back panel	3	Lay down side panel
3	Lay down bottom panel		
3	Lay down front panel		
3	Lay down leg		
3	Lay down shelf		
3	Lay down table top		
4	Pick up bottom panel	4	Pick up back panel
4	Pick up side panel	4	Pick up table top
4	Pick up front panel		
4	Pick up leg		
4	Pick up pin		
4	Pick up shelf		
5	Push table	5	Push table top

(b) Action labels defined for the compositional action classification task from the IKEA-Assembly dataset.

Table 1: Comparison of the original and compositional tasks for the IKEA-Assembly dataset.