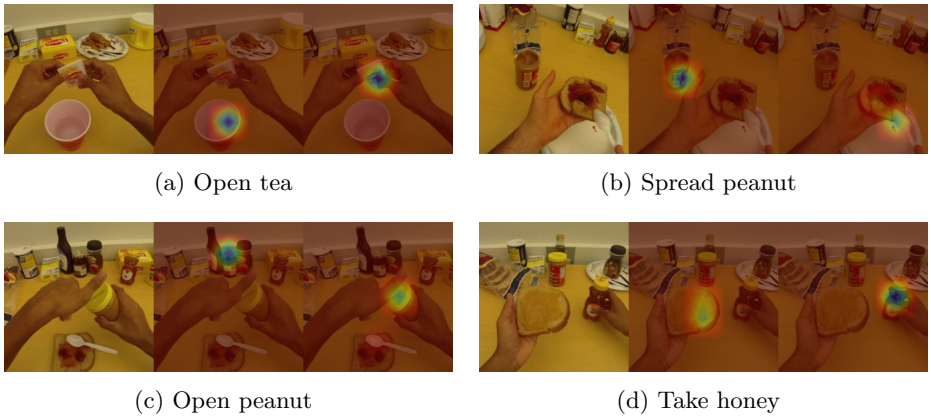
EPIC@ECCV18 Workshop Extended Abstract

Object-centric Attention for Egocentric Activity Recognition

Swathikiran Sudhakaran[1,2] and Oswald Lanz[1]

[1] Fondazione Bruno Kessler, Trento, Italy
[2] University of Trento, Italy
{sudhakaran,lanz}@fbk.eu



(a) Open tea (b) Spread peanut

(c) Open peanut (d) Take honey

Fig. 1: Spatial attention maps obtained for frames from GTEA 61 dataset. First image shows the original frame, second image shows the attention map generated with ResNet-34 trained on imagenet and the last image shows the attention map obtained using our network trained for activity recognition.

In our recent work that will appear at BMVC18 [1] we take on the problem of fine-grained recognition of egocentric activities, which is more challenging than egocentric action recognition. Action recognition [2] involves identifying a generalized motion pattern of hands such as take, put, stir, pour, etc. whereas activity recognition [3] concerns more fine-grained composite patterns such as take bread, take water, put sugar in coffee, put bread in plate, etc. For developing a system capable of recognizing activities, it is pertinent to identify both the hand motion patterns as well as the objects on to which a manipulation is being applied to. Majority of state-of-the-art techniques use hand segmentation [2,3,4], gaze location [3] or object bounding boxes [4] for identifying the location of the relevant objects in the scene that can assist in identifying the activity. These approaches require complex pre-processing which includes human intervention

for generating hand masks or gaze locations of the video frames. Even though there exist wearable devices which can estimate the gaze direction to dispose of the excessive cost of manual annotation, these may cause discomfort to the user or result in inaccuracies during distraction or short interruption of the activity or if the user is wearing glasses.

Considering the aforementioned problems that prevent from leveraging massive collections of natural activity videos, it is essential to develop techniques capable of identifying the relevant objects without being trained with full supervision. Towards this end, we present a CNN-RNN architecture that is trained in a weak supervision setting to predict the raw video-level activity-class label associated with the clip. Our CNN backbone is pretrained for generic image recognition and augmented on top with an attention mechanism that uses class activation maps for spatially selective feature extraction. The memory tensor of a convolutional LSTM then tracks the discriminative frame-based features distilled from the video for activity classification. Our design choices are grounded to fine grained activity recognition because:

- Frame-based activation maps are not bound to reflect image recognition classes, they develop their own representation classes implicitly while training the video-level classification;

- Convolutional LSTM maintains the spatial structure of the input sequence all the way up to the final video descriptor used by the activity classification layer, thus facilitating the spatio-temporal encoding of objects and their locations into the descriptor as they develop into the activity over time.

Our network is therefore able to learn highly specialized attention maps for each frame as can be seen in Figure 1. Our model is trained in a weakly supervised setting using raw video-level activity-class labels. Nonetheless, on standard egocentric activity benchmarks our model surpasses by up to +6% points recognition accuracy the currently best performing method that leverages hand segmentation and object location strong supervision for training.

Our experimental results are summarized in Table 1. We compare the proposed method with state-of-the-art techniques on Georgia Tech Egocentric Activity Datasets with up to 106 different activity classes. The first block in the table shows methods specifically proposed for egocentric activity recognition and the second block shows methods proposed for third person action recognition. The method proposed by Li *et al.* [3] uses hand segmentation for detecting the location of the objects being handled, while Ma *et al.* [4] trains a network for hand segmentation and object localization for obtaining explicit information about the objects and their location. In the proposed method, we make use of the prior knowledge inflated in the network to identify the location of the object that is relevant in identifying the activity class. Figure 1 shows the attention map generated by our network. We can see that the network is capable of identifying the location of the relevant objects present in the frames even though it is trained with weak supervision using the activity-class label. From Table 1, we can see that our method enjoys a notable performance boost, which validates

Methods	GTEA 61*	GTEA 61	GTEA 71	EGTEA
Li *et al.* [3]**	66.8	64	62.1	–
Ma *et al.* [4]**	75.08	73.02	73.24	–
Two stream [5]	57.64	51.58	49.65	41.84
I3D [6]	–	–	–	51.68
TSN [7]	67.76	69.33	67.23	55.93
EGO-RNN(ours)	**77.59**	**79**	**77**	**60.76**

Table 1: Comparison with state-of-the-art methods on popular egocentric datasets, we report recognition accuracy in %. (*: fixed split; **: trained with strong supervision). We release an implementation of our method in pytorch at `github.com/swathikirans/ego-rnn`.

its efficacy in performing egocentric activity recognition. Another point worth mentioning is that the state-of-the-art-techniques for action recognition from third person videos, listed in Table 1, are performing sub-par compared to the methods proposed for egocentric videos. This shows that these methods cannot be considered as standardized techniques for action recognition from videos in general and the importance of developing methods tailored for egocentric videos.

Ongoing work involves exploring the possibility of adding temporal attention since not all frames present in a video are equally representative of the concerned activity. We will further evaluate our method on the recently introduced EPIC-Kitchens dataset [8] and on video recognition problems from third-person views.

This extended abstract excerpts from our forthcoming BMVC18 paper [1] where all the details about the model and the experimental results can be found.

References

1. Sudhakaran, S., Lanz, O.: Attention is All We Need: Nailing Down Object-centric Attention for Egocentric Activity Recognition. In: Proc. BMVC (2018)
2. Singh, S., Arora, C., Jawahar, C.: First Person Action Recognition Using Deep Learned Descriptors. In: Proc. CVPR (2016)
3. Li, Y., Ye, Z., Rehg, J.M.: Delving into Egocentric Actions. In: Proc. CVPR (2015)
4. Ma, M., Fan, H., Kitani, K.M.: Going deeper into first-person activity recognition. In: Proc. CVPR (2016)
5. Simonyan, K., Zisserman, A.: Two-Stream Convolutional Networks for Action Recognition in Videos. In: Proc. NIPS (2014)
6. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proc. CVPR (2017)
7. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In: Proc. ECCV (2016)
8. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. In: Proc. ECCV (2018)