When will you do what? - Anticipating Temporal Occurrences of Activities (Extended Abstract)

Yazan Abu Farha, Alexander Richard, Juergen Gall
University of Bonn, Germany
{abufarha,richard,gall}@iai.uni-bonn.de

Video understanding has gained an increased attention recently. Methods for classifying and segmenting activities in videos or anticipating the very recent future have been very successful. However, such methods are not sufficient for applications where a moving system has to react or interact with humans. Collaborative robots, for instance, have to anticipate the activities of a human in the future for better interactions. In contrast to previous works that anticipate the next activity only within a short time horizon [6, 1, 3], we address the problem of anticipating all activities that will be happening within a time horizon of up to 5 minutes. This includes the classes and order of the activities that will occur as well as when each activity will start and end.

To address this problem, we propose two novel approaches. In both cases, frame-wise action labels are inferred for the observed part using an RNN-HMM [4], then the inferred actions are used to predict the future actions as shown in Figure 1. The first model builds on a recurrent neural network and predicts the future recursively, while the second model builds on a convolutional neural network and predicts future actions in one pass.

The RNN approach receives a sequence of observed action segments as input, where each segment is represented by a 1-hot encoding of the class label and a normalized segment length. Sequentially forwarding all those segments, the RNN predicts the remaining length of the last observed segment as well as a label and a length for the next segment. This prediction is concatenated with the observed segments to form a new input for the network and produce the next prediction. By recursively repeating this process, new predictions are generated until the desired amount of frames is predicted. As loss for a single training example, we use

$$\mathcal{L} = -\log \hat{p}_c + (l_r - \hat{l}_r)^2 + (l_n - \hat{l}_n)^2, \qquad (1)$$

where $\hat{l}_r$ denotes the predicted remaining length of the current action, $\hat{l}_n$ denotes the predicted length of the next action, and $\hat{p}_c$ the predicted class probability of the next action. For training, we minimize the loss, which is summed over all training examples, by backpropagation through time.

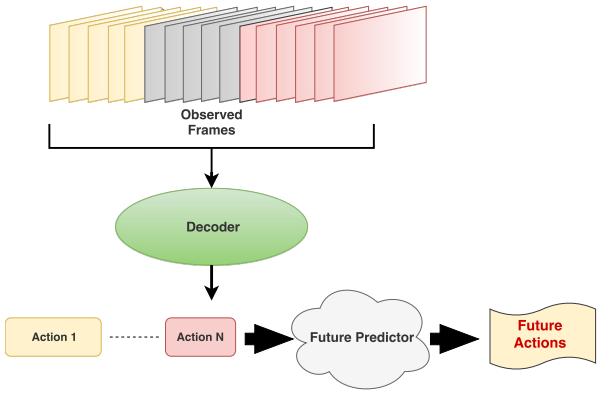


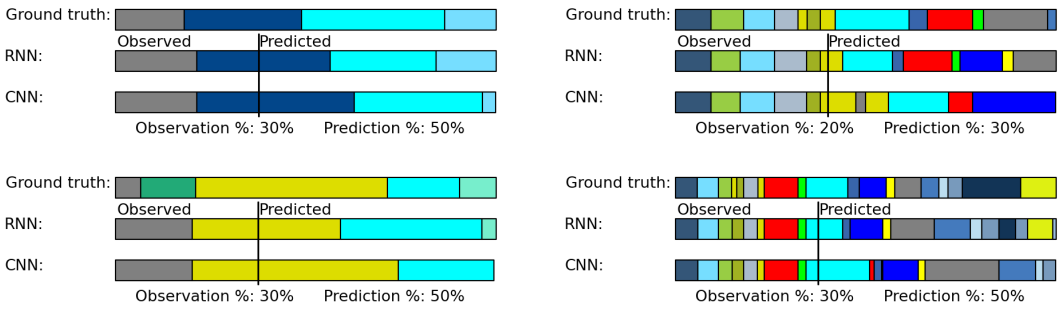
Figure 1. Proposed approach for future action prediction. From the observed frames $\mathbf{x}_1^t$, action labels are inferred by a decoder. The future predictor predicts from the inferred frame labels $\mathbf{c}_1^t$ the labels $\mathbf{c}_{t+1}^T$ that are yet to come.

The second approach builds on a convolutional neural network (CNN). The sequence of inferred activities is converted into a matrix X with C columns and S rows that encodes both the length and the action label information. While the columns correspond to the C action classes, the rows correspond to action segments. Given t observed frames, the number of rows for an action segment of length l is given by $\lfloor \frac{l}{t} S \rfloor$ and, for each row s, $X_{sc} = 1$ for the label c of the corresponding action segment and zero otherwise. By forwarding the matrix X through the CNN, we predict a matrix Y that encodes the length and the action labels of the anticipated activities. In contrast to the RNN approach, the CNN approach anticipates all activities in one pass. To train the network, we use the squared error criterion over all output elements

$$\mathcal{L} = \frac{1}{SC} \sum_{s,c} (Y_{sc} - \hat{Y}_{sc})^2, \qquad (2)$$

where $\hat{Y}$ is the prediction of the network.

We evaluated the two approaches on two challenging datasets that contain long sequences and large variations: the Breakfast dataset [2] and 50Salads [5]. Both approaches



(a) Results on Breakfast (b) Results on 50Salads

Figure 2. Qualitative results for the future action prediction task for both, RNN and CNN without ground-truth observations.

outperform by a large margin two baselines, a grammar based baseline and a nearest neighbor baseline, and also outperform the method of [6] in predicting the immediate future. Figure 2 shows a few example predictions. Both the RNN and the CNN are able to generate accurate predictions that scale well along different videos with varying lengths and huge variations in the possible future actions.

References

[1] J. Gao, Z. Yang, and R. Nevatia. RED: reinforced encoder-decoder networks for action anticipation. In *British Machine Vision Conference (BMVC)*, 2017. 1

[2] H. Kuehne, A. Arslan, and T. Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 780–787, 2014. 1

[3] T. Mahmud, M. Hasan, and A. K. Roy-Chowdhury. Joint prediction of activity labels and starting times in untrimmed videos. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[4] A. Richard, H. Kuehne, and J. Gall. Weakly supervised action learning with RNN based fine-to-coarse modeling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[5] S. Stein and S. J. McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2013. 1

[6] C. Vondrick, H. Pirsiavash, and A. Torralba. Anticipating visual representations from unlabeled video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 98–106, 2016. 1, 2