Supplementary: Temporal Aggregate Representations for Long-Range Video Understanding

Fadime Sener[1,2], Dipika Singhania[2], and Angela Yao[2]

[1] University of Bonn, Germany
[2] National University of Singapore
{sener@cs.uni-bonn.de},{dipika16,ayao}@comp.nus.edu.sg

1 More on Datasets and Features

We provide more statistics about the datasets used in our paper to show a broader comparison about their scale and label granularity.

The Breakfast Actions dataset [7] contains 1712 videos of 10 high level tasks like “making coffee”, “making tea” and so on. There are in total 48 different actions, such as “pouring water” or “stirring coffee”, with on average 6 actions per video. The average duration of the videos is 2.3 minutes. There are 4 splits and we report our results averaged over them. We use two types of frame-wise features: Fisher vectors computed as in [1] and I3D features [2].
The 50Salads dataset [9] includes 50 videos and 17 different actions for a single task, namely making mixed salads. When training on this dataset, we therefore omit task prediction in our model. On average, 50Salads has 20 actions per video due to repetitions. The average video duration is 6.4 minutes. There are 5 splits, and we again average our results over them. We represent the frames using Fisher vectors as in [1].
The EPIC-Kitchens dataset [3] is a large first-person video dataset which contains 432 sequences and 39,594 action segments recorded by participants performing non-scripted daily activities in their kitchen. The average duration of the videos is 7.6 minutes ranging from 1 minute to 55 minutes. An action is defined as a combination of a verb and a noun, e.g. “boil milk“. There are in total 125 verbs, 351 nouns and 2513 actions. The dataset provides a training and test set which contains 272 and 160 videos, respectively. The test set is divided into two splits: Seen Kitchens (S1) where sequences from the same environment are in the training data, and Unseen Kitchens (S2) where complete sequences of some participants are held out for testing. The labels for the test set are not shared, as there is an action anticipation challenge[3] and action recognition challenge[4]. We use the RGB, optical flow and object-based features provided by Furnari and Farinella *et al.* [5]. The minimum and maximum snippet durations,

---

[3] `https://competitions.codalab.org/competitions/20071`
[4] `https://competitions.codalab.org/competitions/20115`

over which we apply pooling, are 0.4s and 115.3s for 50Salads, 0.1s and 64.5s for Breakfast, and 1.2s and 3.0s for EPIC.

Implementation Details : We train our model using the Adam optimizer [6] with batch size 10, learning rate 10^{-4} and dropout rate 0.3. We train for 25 epochs and decrease the learning rate by a factor of 10 every 10th epoch. We use 1024 dimensions for all non-classification linear layers for the Breakfast Actions and 50Salads datasets and 512 dimensions for the EPIC-Kitchens dataset. The LSTMs in dense anticipation have one layer and 512 hidden units. We use intervals of 20 seconds for Breakfast and 50Salads for discretizing the durations in dense anticipation.

2 Model Validation

	cereal	coffee	f.egg	juice	milk	panc.	salat	sand.	s.egg	tea	mean±std
TM	77.8	50.8	57.2	57.2	40.1	39.6	57.9	52.4	59.4	54.2	54.6±10.8
LUT	57.5	**59.9**	56.2	58.8	56.1	57.3	55.1	49.6	**61.2**	60.1	57.2±3.1
LSTM	**79.8**	47.2	52.9	61.2	72.7	**73.9**	**64.3**	46.9	60.5	**68.7**	62.8±11.3
ours	69.8	54.7	**62.5**	**65.7**	**72.9**	66.2	63.6	**64.6**	58.0	64.1	**64.2**±5.2

Table 1. Model validation using GT labels for next action anticipation on the Breakfast Actions, presented are accuracies. We compare transition matrices (TM), lookup tables (LUT), LSTMs, and our temporal aggregates model (without complex activity prediction).

For validating our method's capabilities in modelling sequences, we make baseline comparisons. The simplest approach for solving the next action anticipation task is using a transition matrix (TM) [8], which encodes the transition from one action to the next. A more sophisticated solution is building a lookup table (LUT) of varying length sequences which allows encoding the context in a more explicit manner. The problem with LUTs is that their completeness depends on the coverage of the training data, and they rapidly grow with the number of actions. So far, for next step prediction, RNNs achieve good performance [1], as they learn modelling the sequences.

For our baseline comparisons, instead of frame features, we use the frame-level ground truth labels as input to our model. We compute the TM, LUT and RNN on the ground truth segment-level labels. In Table 1 we present comparisons on the Breakfast Actions for the next action anticipation per complex activity. Overall, transition matrices provide the worst results. LUTs improve the results, as they incorporate more contextual information. Both the RNN and our method outperform the other alternatives, while our method still performs better than the RNN on average. However, applying RNNs requires parsing the past into action sequences [1], which turns the problem into separate segmentation and prediction phases. Our model, on the other hand, can be trained end-to-end, and can represent the long-range observations good enough to outperform RNNs. We

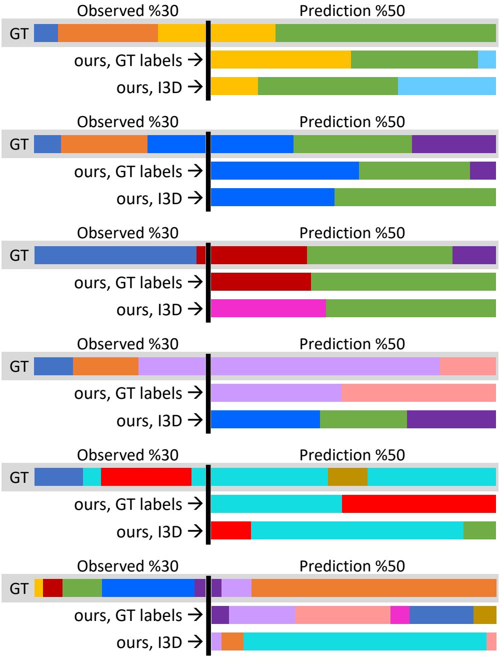

Fig. 1. Qualitative results for dense anticipation on Breakfast Actions dataset when using the GT labels and I3D features. Best viewed in color.

show that our model is doing better than simply learning pairwise statistics of the dataset.

3 Visual Results

In Fig. 1, we provide qualitative results from our method for dense anticipation on the Breakfast Actions dataset. We show our method's predictions after observing 30% of the video. We compare our results when we use the GT labels and I3D features as input.

In Fig. 2, we present qualitative results from our method for next action anticipation on the EPIC-Kitchens dataset for multiple anticipation times τ_α between 0.25 and 2 seconds. We show examples where our method is certain about the next action for all different times. We also show examples where our method's prediction gets more accurate when the anticipation time is closer.

In Fig. 3, we present some visualizations of regions attended by our non-local blocks. We show the five highest weighted spanning snippets (at different granularities). Our model attends different regions over the videos, for example for predicting 'fry egg' when making fried eggs, it attends regions both when pouring oil and cracking eggs. Pouring oil is an important long-range past action

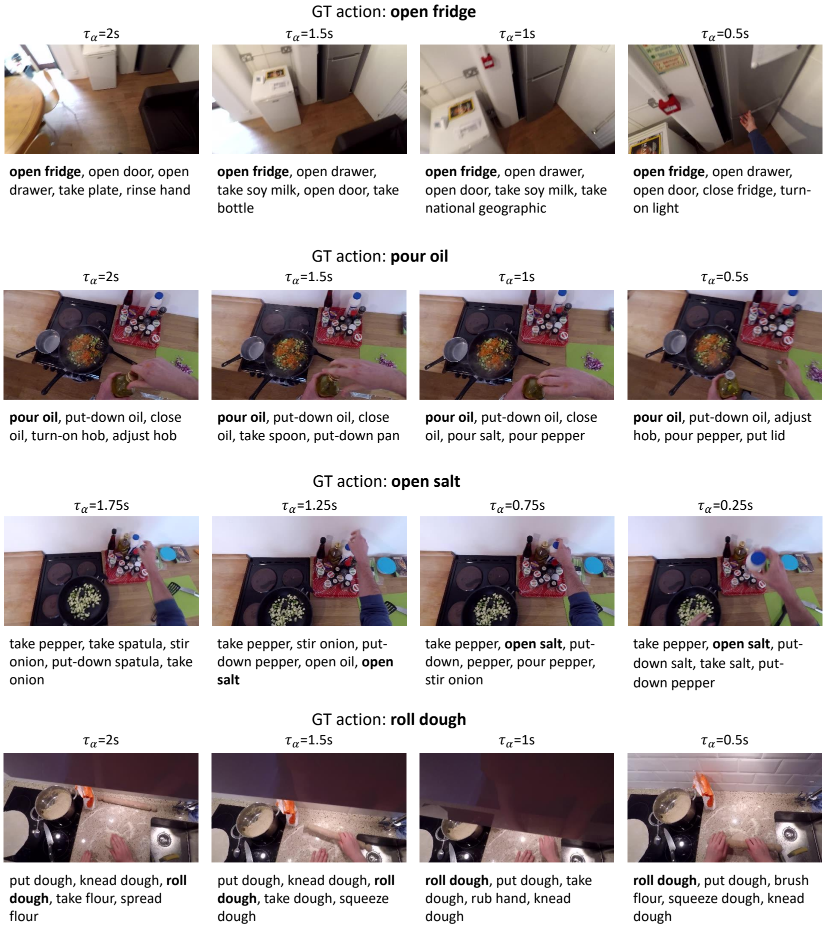
GT action: open fridge
τ_α=2s
τ_α=1.5s
τ_α=1s
τ_α=0.5s
open fridge, open door, open drawer, take plate, rinse hand
open fridge, open drawer, take soy milk, open door, take bottle
open fridge, open drawer, open door, take soy milk, take national geographic
open fridge, open drawer, open door, close fridge, turn-on light
GT action: pour oil
τ_α=2s
τ_α=1.5s
τ_α=1s
τ_α=0.5s
pour oil, put-down oil, close oil, turn-on hob, adjust hob
pour oil, put-down oil, close oil, take spoon, put-down pan
pour oil, put-down oil, close oil, pour salt, pour pepper
pour oil, put-down oil, adjust hob, pour pepper, put lid
GT action: open salt
τ_α=1.75s
τ_α=1.25s
τ_α=0.75s
τ_α=0.25s
take pepper, take spatula, stir onion, put-down spatula, take onion
take pepper, stir onion, put-down pepper, open oil, open salt
take pepper, open salt, put-down, pepper, pour pepper, stir onion
take pepper, open salt, put-down salt, take salt, put-down pepper
GT action: roll dough
τ_α=2s
τ_α=1.5s
τ_α=1s
τ_α=0.5s
put dough, knead dough, roll dough, take flour, spread flour
put dough, knead dough, roll dough, take dough, squeeze dough
roll dough, put dough, take dough, rub hand, knead dough
roll dough, put dough, brush flour, squeeze dough, knead dough

Fig. 2. Exemplary qualitative results for next action anticipation on EPIC-Kitchens dataset, showing the success of our method. We list our Top-5 predictions at different anticipation times, τ_α. The closer we are the better are our model's predictions. Best viewed in color.

for frying eggs. Our method can encode long video durations while attending to salient snippets.

4 Action Anticipation on EPIC-Kitchens

Furnari and Farinella [5] reports prediction results at multiple anticipation times (τ_α) between 0.25s and 2s on EPIC. We compare in Table 2 on the validation set and note that our prediction scores are better than [5] for all time points. Our improvements are greater when the anticipation time decreases.

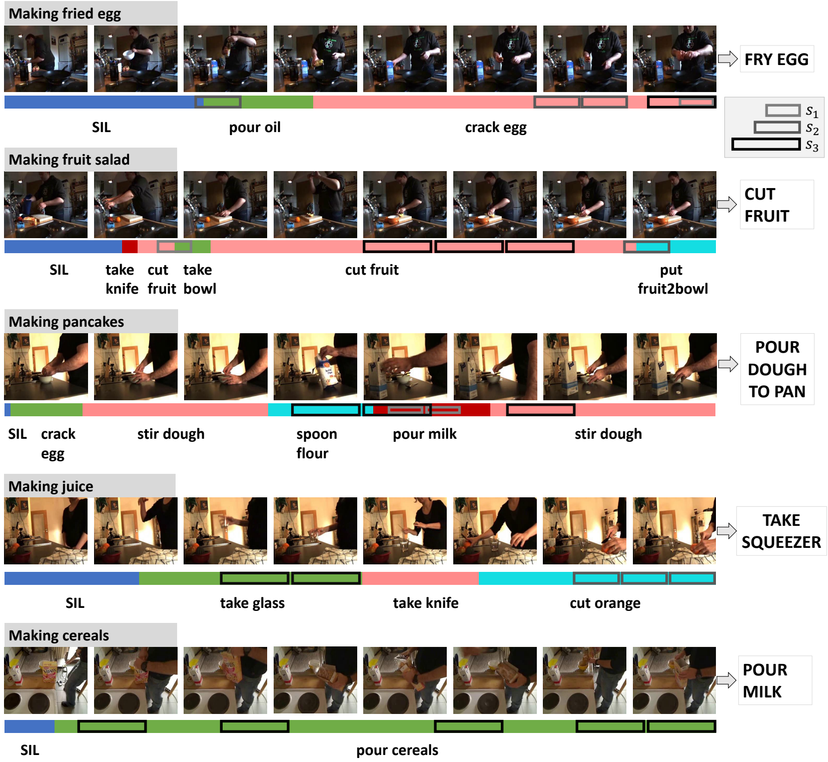

Fig. 3. Attention visualization on the Breakfast Actions dataset for next action anticipation. Rectangles are the top 5 five spanning snippets (at different granularities where K = 10,15,20), weighted highest by the attention mechanism in the Non-Local Blocks (NLB). Best viewed in color.

We report our results for hold-out test data on EPIC-Kitchens Egocentric Action Anticipation Challenge (2020) in Table 3 for seen kitchens (S1) with the same environments as in the training data and unseen kitchens (S2) of held out environments. The official ranking on the challenge is based on the Top-1 action accuracy. Our submission (Team "NUS_CVML") is ranked first on S1 and third on S2 sets. We refer the reader to EPIC-Kitchens 2020 Challenges Report [4] for details on the competing methods.

5 Action Recognition Challenge on EPIC-Kitchens

We present our results for the EPIC-Kitchens Egocentric Action Recognition Challenge 2020 in Table 4 for S1 and S2. Our team "NUS_CVML" is ranked second on S1 and third on S2 sets. Please see EPIC-Kitchens 2020 Challenges Report [4] for further details.

Top-5 ACTION Accuracy%								
τ_a	2s	1.75s	1.5s	1.25s	1.0s	0.75s	0.5s	0.25s
RU [5]	29.4	30.7	32.2	33.4	35.3	36.3	37.4	39.0
ours	**30.9**	**31.8**	**33.7**	**35.1**	**36.4**	**37.2**	**39.5**	**41.3**

Table 2. Action anticipation on EPIC validation set at different anticipation times.

	Top-1 Accuracy%			Top-5 Accuracy%			Precision (%)			Recall (%)		
	Verb	Noun	Action	Verb	Noun	Action	Verb	Noun	Action	Verb	Noun	Action
1st (S1)	37.87	24.10	**16.64**	79.74	53.98	36.06	36.41	25.20	9.64	15.67	22.01	10.05
3rd (S2)	29.50	16.52	**10.04**	70.13	37.83	23.42	20.43	12.95	4.92	8.03	12.84	6.26

Table 3. Action anticipation on EPIC tests sets, seen (S1) and unseen (S2)

	Top-1 Accuracy%			Top-5 Accuracy%			Precision (%)			Recall (%)		
	Verb	Noun	Action	Verb	Noun	Action	Verb	Noun	Action	Verb	Noun	Action
2nd (S1)	66.56	49.60	**41.59**	90.10	77.03	64.11	59.43	45.62	25.37	41.65	46.25	26.98
3rd (S2)	54.56	33.46	**26.97**	80.40	60.98	46.43	33.60	30.54	14.99	25.28	28.39	17.97

Table 4. Action recognition on EPIC tests sets, seen (S1) and unseen (S2)

References

1. Abu Farha, Y., Richard, A., Gall, J.: When will you do what?-anticipating temporal occurrences of activities. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5343–5352 (2018)
2. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6299–6308 (2017)
3. 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 dataset. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. pp. 753–771. Springer International Publishing, Cham (2018)
4. Damen, D., Kazakos, E., Will Price, J.M., Doughty, H.: Epic-kitchens-55 - 2020 challenges report. https://epic-kitchens.github.io/Reports/EPIC-KITCHENS-Challenges-2020-Report.pdf (2020)
5. Furnari, A., Farinella, G.M.: What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In: International Conference on Computer Vision (ICCV) (2019)
6. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
7. Kuehne, H., Arslan, A., Serre, T.: The language of actions: Recovering the syntax and semantics of goal-directed human activities. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). pp. 780–787 (2014)
8. Miech, A., Laptev, I., Sivic, J., Wang, H., Torresani, L., Tran, D.: Leveraging the present to anticipate the future in videos. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 0–0 (2019)
9. Stein, S., McKenna, S.J.: Combining embedded accelerometers with computer vision for recognizing food preparation activities. In: Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing. pp. 729–738. ACM (2013)