

Learning to Abstract and Predict Human Actions - Supplementary Materials

Romero Morais
ralmeidabarata@deakin.edu.au

Vuong Le
vuong.le@deakin.edu.au

Truyen Tran
truyen.tran@deakin.edu.au

Svetha Venkatesh
svetha.venkatesh@deakin.edu.au

Applied Artificial Intelligence Institute
Deakin University
Australia

1 Introduction

In this supplementary material package we include:

1. Statistics and other additional information about our *Hierarchical Breakfast* annotations.
2. Additional quantitative and qualitative evaluation results.

Code/data is available at github.com/RomeroBarata/hierarchical_action_prediction.

2 Hierarchical Breakfast Annotation Analysis

We annotated 1717 videos into a two-level hierarchy: coarse activities and fine actions. This resulted in 25537 annotated segments, with 6549 of them being coarse activities and 18988 of them being fine actions. At the end of the annotation, there were 30 unique coarse activities and 140 unique fine actions annotated across the whole dataset.

In Fig. 1 we can see the number of times each coarse activity got annotated. In Fig. 2 we can see the number of times the top 30 fine actions were annotated (we show the top 30 to avoid clutter). Some activities are not frequent, since the preparation of breakfast meals can widely vary from person to person. For instance, not everyone add sugar to their coffee. These variations in behavior are natural and were all annotated.

3 Additional Results

3.1 Hierarchical Breakfast Dataset

The F1@0.25 values for the results in Fig. 3 of the main paper are shown here on Table 1.

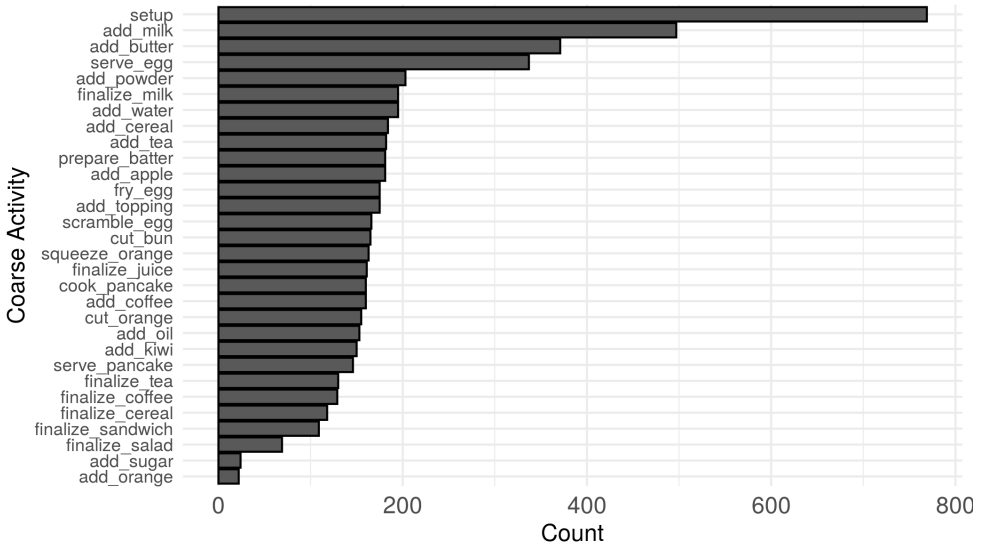


Figure 1: Distribution of the annotated coarse activities for the Hierarchical Breakfast Actions dataset. Coarse activity name is shown on the y-axis whereas the number of times the activity appeared is shown on the x-axis.

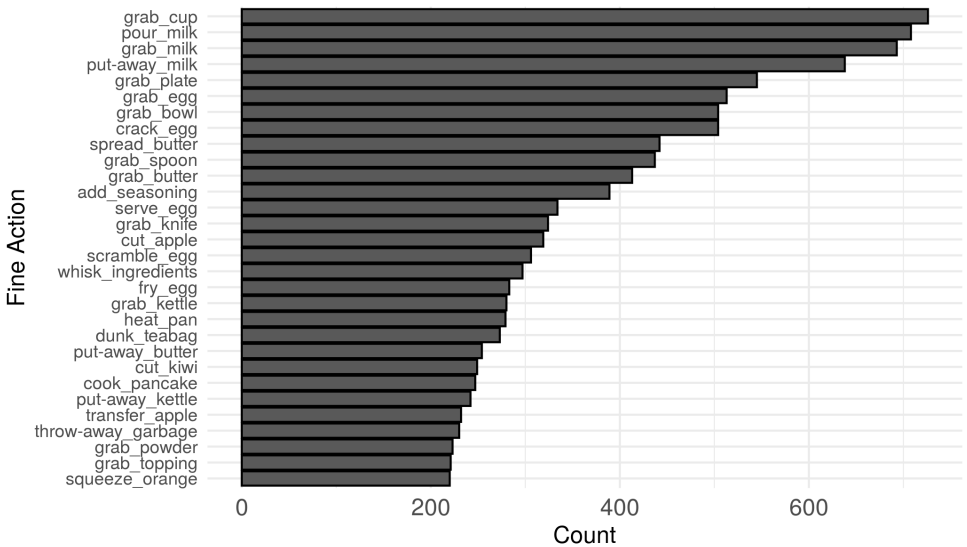


Figure 2: Distribution of the top-30 annotated fine actions for the Hierarchical Breakfast Actions dataset. We show here only the top-30 fine actions to avoid clutter. Fine action name is shown on the y-axis whereas the number of times the action appeared is shown on the x-axis.

Table 1: F1@0.25 of HERA and related methods on the Hierarchical Breakfast Actions dataset. For this experiment, the methods are allowed to observe a percentage of the video (20% or 30%) and need to predict the whole unseen future (70% or 80%). The results are an average of a 4-fold cross-validation and higher results are better.

Observe		20%						30%				
Predict		10%	20%	30%	50%	70%	80%	10%	20%	30%	50%	70%
Coarse	Dummy	87.3%	76.9%	68.1%	53.7%	42.6%	36.4%	88.0%	77.5%	69.9%	55.4%	40.1%
	Baseline 0	80.9%	72.8%	67.1%	64.9%	62.1%	66.9%	76.9%	69.2%	67.9%	65.6%	67.7%
	Baseline 1	71.1%	65.9%	63.8%	63.4%	61.3%	65.7%	62.7%	60.0%	58.8%	58.1%	61.0%
	Baseline 2	76.9%	69.8%	63.1%	59.4%	58.9%	60.5%	70.9%	63.5%	61.7%	61.2%	61.4%
	Farha <i>et al.</i> [10]	84.2%	79.1%	76.1%	75.8%	71.8%	74.0%	85.6%	79.3%	77.7%	76.2%	76.1%
	Farha2 <i>et al.</i> [10]	87.4%	82.1%	76.0%	70.4%	65.8%	65.4%	87.0%	79.2%	75.3%	70.6%	66.8%
Fine	Dummy	62.2%	41.8%	29.6%	16.7%	10.3%	7.5%	66.0%	46.4%	35.4%	21.8%	12.1%
	Baseline 0	36.2%	27.1%	25.0%	21.7%	21.4%	21.4%	35.1%	27.0%	24.8%	22.4%	22.2%
	Baseline 1	42.5%	32.0%	27.9%	25.8%	24.8%	25.3%	38.5%	30.4%	28.1%	26.5%	26.9%
	Baseline 2	38.9%	28.5%	24.9%	22.8%	22.7%	22.6%	39.9%	31.5%	28.9%	27.5%	26.0%
	Farha <i>et al.</i> [10]	63.7%	52.9%	44.5%	37.1%	32.1%	30.4%	66.1%	54.3%	47.8%	40.3%	33.1%
	Farha2 <i>et al.</i> [10]	62.7%	54.2%	48.1%	39.8%	34.8%	32.5%	66.7%	55.3%	48.7%	40.5%	33.7%
Coarse	HERA	86.2%	80.7%	76.8%	76.9%	70.9%	73.9%	88.2%	81.8%	81.4%	78.4%	78.1%
Fine		65.3%	54.0%	47.1%	39.8%	34.9%	34.3%	69.3%	56.5%	48.9%	41.5%	37.6%

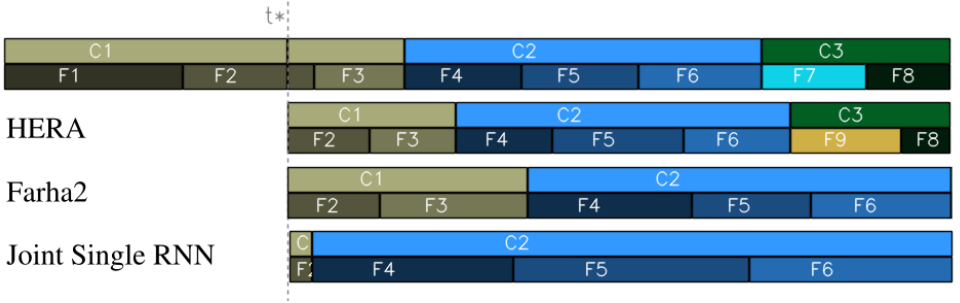


Figure 3: Qualitative evaluation of predictions of different methods on task “prepare cereal”. The first timeline shows the observed and ground-truth future. Others show future predictions of corresponding methods. C1: add cereal, C2: add milk, C3: finalize cereal; F1: grab cereal, F2: pour cereal, F3: put away cereal, F4: grab milk, F5: pour milk, F6: put away milk, F7: stir cereal, F8: grab bowl, F9: grab spoon.

Additional qualitative results are shown in Figs. 3 and 4. In these two examples, we can see that in the short-term both HERA and Farha2 make predictions well aligned with the ground-truth (e.g. F2 and F3 in Fig. 3), but as we move towards long-term predictions mistakes made by Farha2 in the fine-level quickly accumulate and generate misaligned predictions. In Fig. 3, for instance, F4 was too long and from this point on Farha2 predictions F5 and F6 completely misaligned with the ground-truth. HERA, on the other hand had more success in correctly aligning the predicted fine actions with the ground-truth since these predictions built on successful predictions at the coarse level.

3.2 50 Salads Dataset

The F1@0.25 attained by HERA and related methods are shown on Table 2.

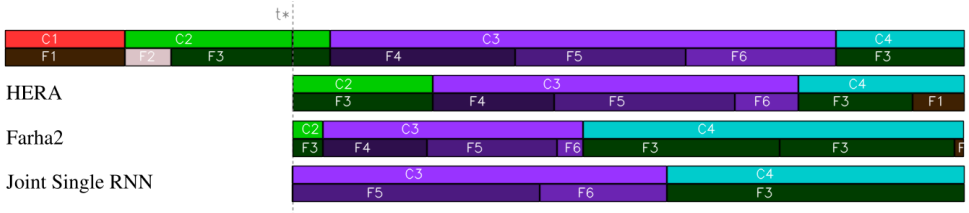


Figure 4: Qualitative evaluation of predictions of different methods on task “prepare cereal”. The first timeline shows the observed and ground-truth future. Others show future predictions of corresponding methods. C1: setup, C2: add tea, C3: add water, C4: finalize tea; F1: grab cup, F2: grab teabag, F3: dunk teabag, F4: grab kettle, F5: pour water, F6: put away kettle.

Table 2: F1@**0.25** of HERA and related methods on the mid and fine levels of the 50 Salads dataset. For this experiment, the methods are allowed to observe a percentage of the video (20% or 30%) and need to predict the whole unseen future (70% or 80%). The results are an average of a 5-fold cross-validation and higher results are better.

Observe		20%						30%				
		Predict	10%	20%	30%	50%	70%	80%	10%	20%	30%	50%
Mid	Dummy	46.2%	23.5%	12.3%	1.1%	0.0%	0.0%	49.1%	23.3%	14.0%	3.1%	0.4%
	Independent Single RNN	32.3%	23.1%	15.3%	8.9%	6.8%	6.3%	37.5%	25.0%	20.0%	12.3%	8.3%
	Joint Single RNN	40.3%	25.3%	20.8%	13.4%	8.4%	7.6%	43.7%	25.6%	21.0%	13.5%	8.0%
	Synced Pair RNN	41.4%	25.8%	19.9%	13.4%	8.5%	7.8%	41.6%	28.6%	20.5%	13.0%	7.9%
	Farha <i>et al.</i> [10]	55.7%	41.7%	35.3%	29.7%	26.8%	28.2%	46.8%	33.8%	27.0%	22.1%	22.9%
Fine	Dummy	19.2%	4.8%	1.3%	0.5%	0.0%	0.0%	18.8%	5.0%	1.9%	0.2%	0.0%
	Independent Single RNN	21.1%	11.8%	8.8%	5.6%	3.8%	3.3%	18.4%	8.9%	5.7%	3.9%	2.3%
	Joint Single RNN	15.8%	7.4%	5.8%	3.6%	2.4%	2.1%	14.6%	8.3%	6.5%	3.8%	2.2%
	Synced Pair RNN	22.1%	10.4%	7.0%	4.6%	2.6%	2.3%	20.5%	8.8%	5.4%	3.2%	1.7%
	Farha <i>et al.</i> [10]	24.8%	17.7%	14.4%	9.8%	7.5%	7.8%	29.7%	19.1%	13.8%	8.7%	7.5%
Mid	HERA	46.8%	34.1%	24.8%	18.9%	15.7%	19.9%	41.5%	31.3%	23.7%	16.3%	18.9%
Fine		21.9%	13.1%	9.0%	5.9%	5.3%	8.3%	20.5%	12.5%	9.7%	6.4%	8.7%

References

- [1] Yazan Abu Farha, Alexander Richard, and Juergen Gall. When will you do what? - anticipating temporal occurrences of activities. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5343–5352. IEEE, June 2018. doi: 10.1109/CVPR.2018.00560.