Supplemental - We Have So Much In Common: Modeling Semantic Relational Set Abstractions in Videos

Alex Andonian*[1], Camilo Fosco*[1], Mathew Monfort[1], Allen Lee[1], Rogerio Feris[2], Carl Vondrick[3], and Aude Oliva[1]

[1] Massachusetts Institute of Technology
[2] MIT-IBM Watson AI Lab
[3] Columbia University

1 Introduction

In this document, we provide more details about the human experiments performed for this paper. We additionally show ablation experiments, qualitative results and quantitative comparisons to further motivate our model.

2 The VidRank and Odd One Out games

As explained in Section 3.1, we created the VidRank game to measure human performance on the Set Completion task. We built a web-based UI to show a set of 2 to 4 *reference* videos along 5 *query* videos. The user was tasked to rank the *query* videos according to how close they were to the general concept represented by the *reference* videos. Figure 1 shows an example question from our game. Quality was ensured through "vigilance" questions, where one query was either obviously similar to the references (same class and similar visual properties) or drastically different. We positioned three vigilance rounds per set of 10 questions, and discarded data from users that failed more than one.

To obtain human performance for the Odd One Out task, we created the Odd One Out game using a similar UI, where users were tasked with finding the video that did not belong to the set. We showed users sets of 3-5 videos, and allowed them to select the suspected outlier. We again positioned "vigilance" rounds, where one video was clearly dissimilar to the rest (all videos shared the same action class while the odd one out was selected from a dissimilar class).

To obtain the baselines presented in the main paper, we collected human responses from our two games through Amazon Mechanical Turk (AMT). We obtained 400 human responses for each task and dataset, ensuring that at least 20 participants attempted each question. We validated the coherence of the responses by computing split-half consistency curves. These curves are produced by separating the users into two groups, calculating scores inside each group separately, and correlating the resulting values. It is a measure of agreement between participants, and a rank correlation of 0.7 or higher generally indicates

high consistency across humans on the evaluated task, and thus good quality data. We show the results in Figure 2.

We measured the rank correlation between human decision-making and our algorithmic ranking for the set completion task, and the accuracy at selecting the video farthest to the abstraction for the odd one out task. We show the results for both datasets in Figure 3.

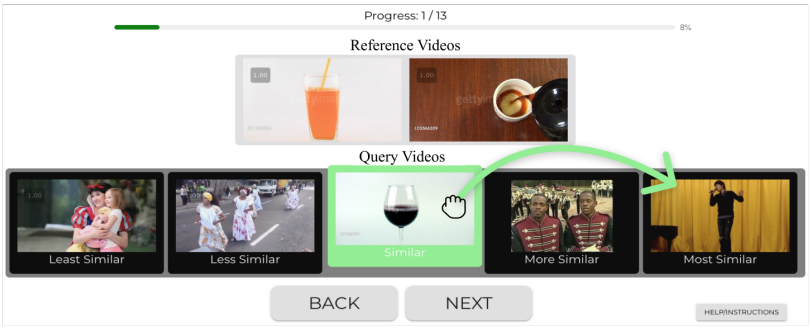

Fig. 1: **VidRank UI** With an intuitive click and drag interface, humans are tasked to rank 5 query videos (bottom row) according to how close they are to the abstract action inferred from the reference videos (top row). As the videos are 3 to 10 sec. long, reference and query videos are shown simultaneously as GIFs. They contain one semantic action for Kinetics, and multiple related ones for M-MiT (e.g. boating, sailing).

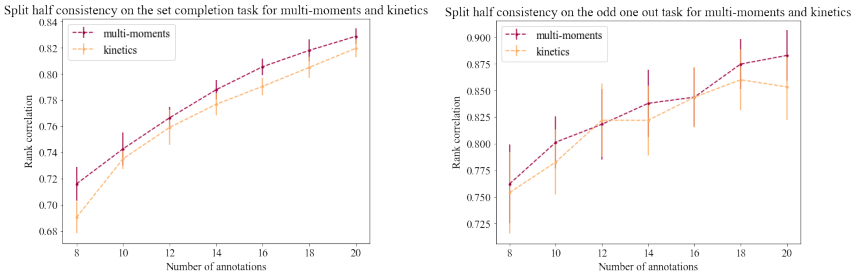

Fig. 2: Split-half consistency curves of human responses for the Set Completion task (left) and Odd One Out task (right). As can be seen, our results show a strong level of consistency between human groups, attaining 0.82 and 0.85 rank correlation for the set completion and odd one out tasks respectively.

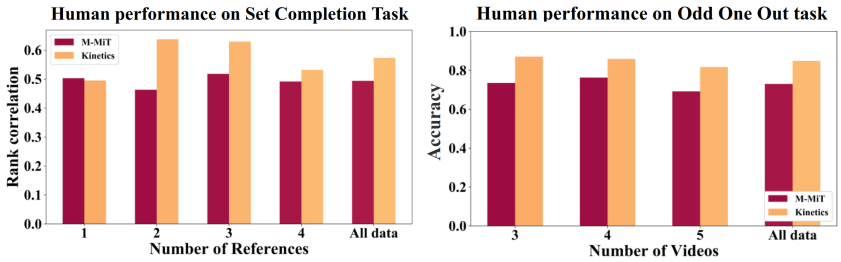

Fig. 3: Human performance on the set completion (left) and odd one out (right) tasks. We show results in terms of rank correlation for the set completion task, and accuracy for the odd one out task. Humans were tested with the same sets that were used for evaluating the model. Our model is very close to human performance.

3 Word Embeddings Ablation Analysis

In section 4.2 of the main paper, we described how we utilize natural language supervision in the form of pretrained word embeddings to generate high-level representations of similarities across a set of videos. In this section, we investigate the contribution of this additional supervision by ablating it and observing how it affects performance on the tasks described in Section 6. For all the referenced tables in this section, the model labeled SAM is the function $g \circ B$ that is trained *without* the embedding loss $\mathcal{L}_{mse}$, and SAM + Embed corresponds to $e \circ g \circ B$ trained with the embedding loss included.

		N = 2		N = 3		N = 4	
Dataset	**Model**	**Top1**	**Top5**	**Top1**	**Top5**	**Top1**	**Top5**
M-MiT	SAM	32.1	64.9	37.2	73.0	42.2	78.3
	SAM + Embeddings	34.0	66.9	41.1	77.1	47.2	83.8
Kinetics	SAM	58.9	85.1	63.7	90.7	67.0	93.8
	SAM + Embeddings	60.5	86.0	65.3	91.6	69.9	94.6

Table 1: **Set Abstraction Ablations:** Evaluation on Classification accuracy (percent) of SAM trained with and without the embedding loss.

Table 1, which lists the set abstraction classification performance of our proposed model with and without the embedding loss $\mathcal{L}_{mse}$, shows that this additional supervision signal does not disrupt performance on the primary training task. In fact, this loss actually helps to improve performance on both datasets, with the Kinetics model showing the larger of the two effects.

While the embedding loss term produced modest improvements to performance on the set abstraction task, the value of the learned semantic representation is significant in downstream tasks. Table 2 makes a similar comparison

Dataset	Model	N = 1	N = 2	N = 3	N = 4	Avg
M-MiT	SAM	0.437	0.491	0.515	0.525	0.492
	SAM + Embeddings	0.471	0.515	0.555	0.558	0.525
Kinetics	SAM	0.354	0.476	0.481	0.497	0.452
	SAM + Embeddings	0.523	0.627	0.659	0.606	0.604

Table 2: **Set Completion Ablation:** Rank Correlation of our model trained with and without the embedding loss term computed using the embedding distance between a video and the abstraction (embedding) of a *reference* set of size N on the *set completion* task.

		N = 3		N = 4	
Dataset	Model	Top-1	Top-2	Top-1	Top-2
M-MiT	SAM	60.90	81.35	62.00	76.35
	SAM + Embeddings	85.90	92.80	83.18	91.44
Kinetics	SAM	64.20	83.00	68.37	76.03
	SAM + Embeddings	85.90	92.80	83.18	91.44

Table 3: **Odd One Out detection ablation:** We investigate the how the embedding loss affect our model's ability to predict the element that does not belong to the set.

between our model trained with and without the embedding loss, revealing that the language-enhanced features significantly outperform those learned to only predict the abstract category. However, it should be noted that even the features used to predict the abstraction show large performance gains over those obtained from the model baseline described in Section 6.3. The odd one out detection task enjoys similar performance gains as a result of using the learned language embeddings, as seen by the trends in Table 3.

4 Additional Quantitative Results

As explained in section 4.1, we experimented with I3D [1] and 3DResNet50 [3] backbones to choose our final video feature network B. We performed our tests on the Kinetics dataset, and we present the results for our three tasks in tables 4, 5 and 6. We observed comparable performance on recognizing set abstractions, but more dominant results from 3DResNet50 on our two other tasks. As our I3D numbers are generally lower than the 3DResNet50 results, we chose the latter network as our main backbone.

Dataset	Model	N = 2		N = 3		N = 4	
		Top1	Top5	Top1	Top5	Top1	Top5
Kinetics	Chance	0.44	2.18	0.44	2.18	0.44	2.18
	3DResNet50	29.9	49.1	22.1	42.8	17.9	40.4
	I3D	27.7	48.1	21.7	40.9	17.1	39.1
	3DResNet50 (BCE)	2.8	25.0	2.2	22.2	0.5	22.5
	I3D (BCE)	3.6	23.9	4.1	24.2	0.9	21.7
	3DResNet50+RN [4]	53.9	83.0	61.6	90.2	66.0	93.8
	I3D+RN [4]	49.3	78.8	63.3	89.3	71.7	94.5
	3DResNet50+SAM (Ours)	**60.5**	**86.0**	**65.3**	**91.6**	69.9	**94.6**
	I3D+SAM (Ours)	54.5	84.8	65.0	91.1	**73.2**	94.5

Table 4: **Recognizing Set Abstractions:** Classification accuracy (percent) of the models evaluated on the set abstraction task. Here, N is the number of elements in the set, and the top$k$ chance level is the sum of the frequency of the top$k$ most frequent abstract nodes presented during evaluation.

Dataset	Model	N = 1	N = 2	N = 3	N = 4	Avg
Kinetics	Human Baseline	0.432	0.653	0.629	0.606	0.58
	3DResNet50	0.339	0.421	0.431	0.459	0.413
	I3D	0.321	0.409	0.412	0.444	0.411
	3DResNet50+RN [4]	-	0.491	0.487	0.489	0.489
	I3D+RN [4]	-	0.485	0.479	0.486	0.485
	3DResNet50+SAM (Ours)	**0.523**	**0.627**	**0.659**	**0.606**	**0.604**
	I3D+SAM (Ours)	0.508	0.607	0.628	0.589	0.595

Table 5: **Set Completion:** Rank Correlation of our model (3DResNet50+SAM), a baseline (3DResNet50) and human ranking to the ranking achieved using the embedding distance between a video and the abstraction of a *reference* set of size N on the *set completion* task.

Dataset	Model	N = 3		N = 4	
		Top-1	Top-2	Top-1	Top-2
Kinetics	Human Baseline	87.31	-	85.40	-
	3DResNet50	65.15	82.65	69.62	81.48
	I3D	63.91	81.12	67.92	81.08
	3DResNet50+RN [4]	40.11	70.97	30.48	54.41
	I3D+RN [4]	39.19	70.97	28.98	52.11
	3DResNet50+O3N [2]	55.14	80.59	66.00	81.80
	I3D+O3N [2]	52.52	78.89	67.33	82.29
	3DResNet50+SAM (Ours)	**85.90**	**92.80**	**83.18**	**91.44**
	I3D+SAM (Ours)	83.71	91.24	83.03	90.85

Table 6: **Odd One Out detection accuracy:** Predict the element that does not belong to the set. The language-enhanced features from the set abstraction network are compared with the features from the corresponding base model.

5 Additional Qualitative results

Recognizing Set Abstractions. In Figure 4, we show all the predicted outputs from our Set Abstraction Model on a single validation example. Given a set of 4 videos, we predict a class for each element in the power set, totalling 15 predictions in this case.

Fig. 4: All predicted outputs from our Set Abstraction Model on a validation example. We predict a class for each element in the power set of the given videos. We show confidence in parenthesis, and ground truth in brackets.

Set Completion. In Figure 5, we show several qualitative examples of the set completion task. We find that our model is able to correctly understand underlying concepts regardless of actors (e.g. man leaping, dog rising in the last row) and match similar concepts such as the fish on the pan with the videos of cooking third row). We observe that although our model is trained with supervision from actions, some perceptual properties seem to be recognized and used to choose the video that completes the set.

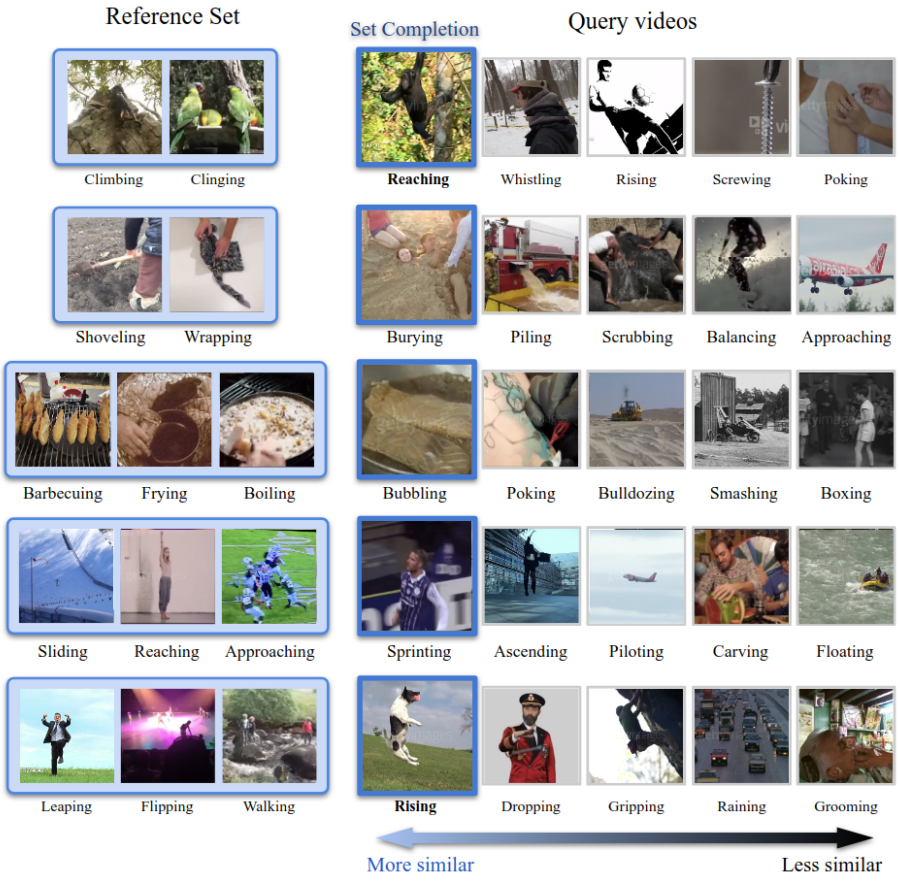

Fig. 5: More qualitative examples on the set completion task. Our model is able to easily align concepts such as burying with actions shoveling and wrapping, or sprinting with exercise related actions.

Finding The Odd One Out. In Figure 6, we show several qualitative examples for the odd one out task with set sizes of 3 and 4, including some failure cases.

Our model seems to fail in cases that could be ambiguous to humans if the name of the action was not given.

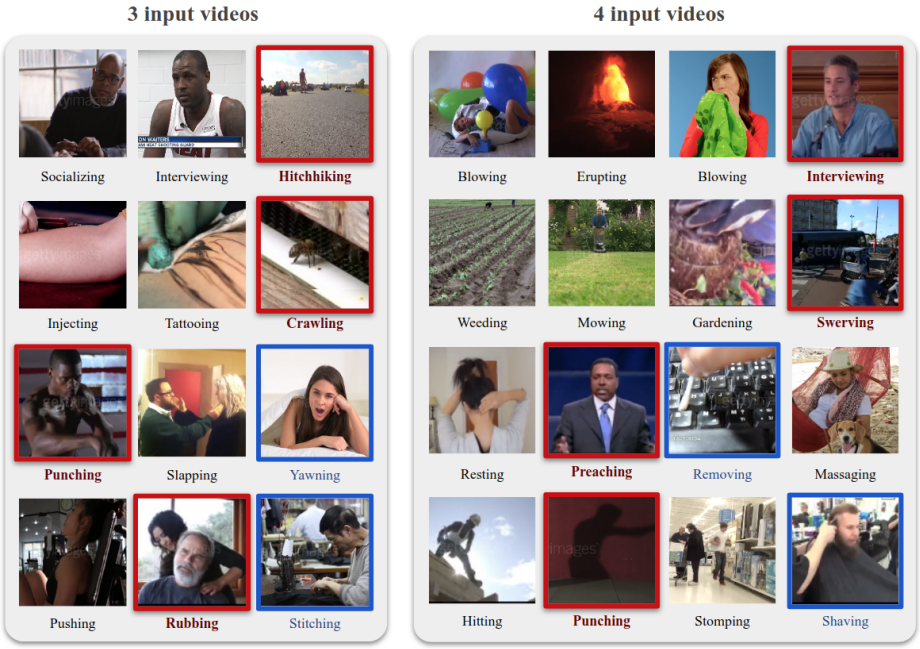

Fig. 6: Qualitative examples on the odd one out task, for 3 (left) and 4 (right) videos in the set We show some failure cases in the last two rows. The ground truth is highlighted in blue, while the model selection is highlighted in red.

References

1. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. Proc. ICCV (2017)
2. Fernando, B., Bilen, H., Gavves, E., Gould, S.: Self-supervised video representation learning with odd-one-out networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
4. Santoro, A., Raposo, D., Barrett, D.G., Malinowski, M., Pascanu, R., Battaglia, P., Lillicrap, T.: A simple neural network module for relational reasoning. In: Advances in neural information processing systems. pp. 4967–4976 (2017)