CLASTER: Clustering with Reinforcement Learning for Zero-Shot Action Recognition Supplementary

Shreyank N Gowda[1], Laura Sevilla-Lara[1], Frank Keller[1], and Marcus Rohrbach[2]

[1] University of Edinburgh
[2] Meta AI

Summary

Section 1 shows how using clustering has a regularization effect on the seen classes with all samples as well as seen classes with a subset of samples.
Section 2 shows that our approach outperforms prior work with a high statistical significance in most scenarios.
Section 3 shows that our approach is consistently outperforming prior work when evaluated on the same split.
Section 4 and Figure 3 shows the performance w.r.t. the number of clusters, which is stable across a wide range of values for this hyper parameter.
Section 5 shows the comparison of aggregation strategies and interaction between visual and semantic features.
Section 6 and Table 4 report the performance of seen and unseen classes separately for the case of GZSL.

1 Regularization Effect of Clustering

In the main paper, we showed the regularization effect that clustering had when using 6, 10 and 51 clusters in comparison to no clusters. Here, we look at the same effect with 20 and 35 clusters as well. We see consistent improvements of over 15% in accuracy for the unseen classes using the proposed CLASTER representation compared to no clustering.

In addition, we show that using only 35% of the data of seen classes for training also benefits from clustering on the unseen classes. This can be seen in Figure 2 While in the case of seen classes, using no clustering has the highest validation accuracy, at test time for the unseen classes, clustering leads to the best results. There are a few interesting points to note here. First, no clustering results in clear overfitting. The training accuracy reaches over 80% while the validation accuracy reaches a peak of 46% before dropping. However, using clustering results in the training and validation curves to be really close to each other. Another interesting point is that when there is a limited number of samples, having more clusters results in better performance at test time. This

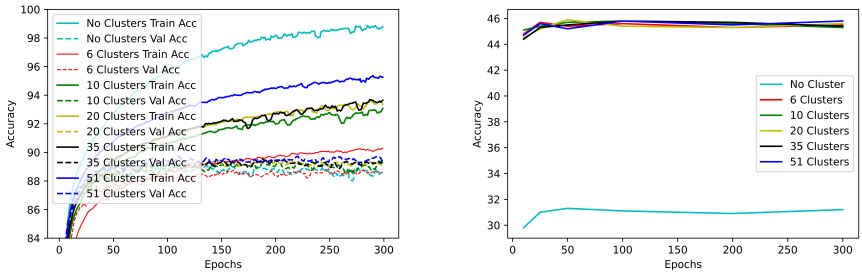

Fig. 1. Left: Learning curve for the seen classes. Right: Accuracy curve for the unseen classes. The clustering-based representation avoids overfitting, which in the case of seen classes means that the gap between validation and training accuracy is smaller than in the vanilla representation. This regularization effect improves the accuracy in unseen classes.

was not the case when we had all samples for the seen classes. When having all samples at training time, the number of clusters resulted in the same average accuracy as can be seen in Section 4.

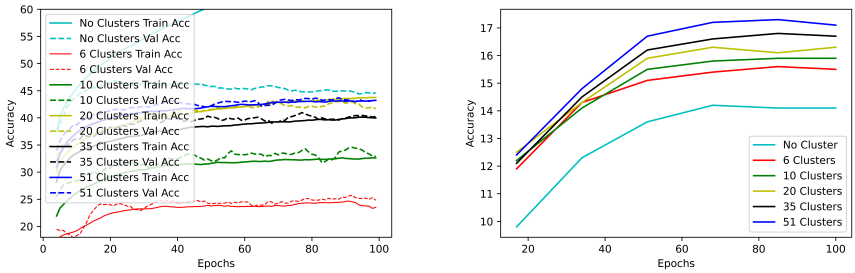

Fig. 2. Left: Learning curve for the seen classes using 35% of the data. Right: Learning curve for the unseen classes. The clustering-based representation avoids overfitting, which in the case of seen classes means that the gap between validation and training accuracy is smaller than in the vanilla representation. This regularization effect improves the accuracy in unseen classes.

2 Statistical Significance

We consider the dependent t-test for paired samples. This test is utilized in the case of dependent samples, in our case different model performances on the same random data split. This is a case of a paired difference test. This is calculated as shown in Eq 1.

$$t = \frac{\bar{X}_D - \mu_0}{s_D/\sqrt{n}} \tag{1}$$

Where $\bar{X}_D$ is the average of the difference between all pairs and s_D is the standard deviation of the difference between all pairs. The constant μ_0 is zero in case we wish to test if the average of the difference is different; n represents the number of samples, $n = 10$ in our case. The comparisons can be seen in Table 1. The lower the value of 'p', higher the significance.

As we can see, our results are statistically significant in comparison to both OD [3] and WGAN [4] in both ZSL and GZSL. We also see that our results are statistically significant for both HMDB51 and Olympics in comparison to E2E [1]. In GZSL, OD [3] also achieves results that are significantly different in comparison to WGAN [4].

Pairs	Dataset	t-value	Statistical significance(p<0.05)	Type
CLASTER and OD [3]	UCF101	-15.77	Significant, p<0.00001	ZSL
CLASTER and WGAN [4]	UCF101	-9.08	Significant, p<0.00001	ZSL
CLASTER and E2E [1]	UCF101	-0.67	Not Significant, p = 0.26	ZSL
OD [3] and WGAN [4]	UCF101	-1.70	Not Significant, p=0.12278	ZSL
CLASTER and OD [3]	HMDB51	-4.33	Significant, p=0.00189	ZSL
CLASTER and WGAN [4]	HMDB51	-5.54	Significant, p=0.00036	ZSL
CLASTER and E2E [1]	HMDB51	-3.77	Significant, p = 0.00219	ZSL
OD [3] and WGAN [4]	HMDB51	-3.71	Significant, p=0.00483	ZSL
CLASTER and OD [3]	Olympics	-9.06	Significant, p<0.00001	ZSL
CLASTER and WGAN [4]	Olympics	-11.73	Significant, p<0.00001	ZSL
CLASTER and E2E [1]	Olympics	-2.72	Significant, p = 0.012	ZSL
OD [3] and WGAN [4]	Olympics	-2.47	Significant, p=0.03547	ZSL
CLASTER and OD [3]	UCF101	-4.51	Significant, p=0.00148	GZSL
CLASTER and WGAN [4]	UCF101	-5.49	Significant, p=0.00039	GZSL
OD [3] and WGAN [4]	UCF101	-3.16	Significant, p=0.01144	GZSL
CLASTER and OD [3]	HMDB51	-5.08	Significant, p=0.00066	GZSL
CLASTER and WGAN [4]	HMDB51	-7.51	Significant, p=0.00004	GZSL
OD [3] and WGAN [4]	HMDB51	-5.27	Significant, p=0.00051	GZSL
CLASTER and OD [3]	Olympics	-5.79	Significant, p=0.00026	GZSL
CLASTER and WGAN [4]	Olympics	-8.39	Significant, p=0.00002	GZSL
OD [3] and WGAN [4]	Olympics	-6.22	Significant, p=0.00014	GZSL

Table 1. Comparison of the t-test for different pairs of models on the same random split. Lower the value of 'p', higher the significance. As we can see, our results are statistically significant in comparison to both OD [3] and WGAN [4] in both ZSL and GZSL. For GZSL, OD [3] also achieves results that are significant in comparison to WGAN [4].

3 Average of Differences in Performance for Same Splits

Since the performance of the model varies for each random split (as witnessed by the standard deviation values), we average the difference in performance between CLASTER, OD, WGAN and E2E on the same splits. We believe that this gives us a better metric to check the performance of CLASTER with the other approaches. The results are depicted in Table 2.

Models	Setting	Olympics	HMDB51	UCF101
Ours and WGAN [4]	ZSL	17.5 ± 4.5	7.0 ±3.8	17.4 ± 5.7
Ours and OD [3]	ZSL	13.6 ± 4.5	2.4 ± 1.6	14.3 ± 2.7
Ours and E2E [1]	ZSL	2.6 ± 2.8	3.7 ± 2.8	0.4 ± 1.8
Ours and WGAN [4]	GZSL	11.2 ± 4.0	9.3 ± 3.7	8.1 ± 4.4
Ours and OD [3]	GZSL	4.6 ± 2.4	5.2 ± 3.1	2.7 ± 1.8

Table 2. Comparing the average of the difference in performance for recent state-of-the-art approaches in zero-shot and generalized zero-shot action recognition on the same splits. All results were computed using sen2vec as the embedding. We can see that we outperform recent approaches in every scenario.

4 Number of Clusters

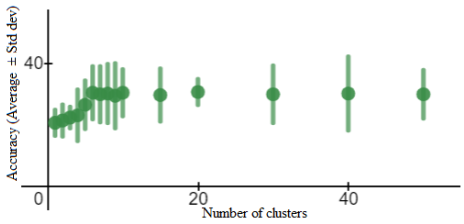

Fig. 3. Effect of using different number of clusters. The green line represents the standard deviation. The reported accuracy is on the UCF101 dataset. As can be seen, the average cluster accuracy increases till about 6 clusters and then remains more or less constant. The vertical lines correspond to the standard deviation.

We test using different number of clusters on the UCF-101 dataset and show the results in Figure 3. These are for 5 runs on random splits. As we can see, the average accuracy increases until 6 clusters, and after that remains more or less constant. Thus, we use 6 clusters and continue with the same number for both HMDB51 and Olympics. For images, similarly, we used 5 random splits of

CUB and found the performance stabilizes after having 9 clusters and use the same number of clusters for the other image datasets.

5 Comparison of aggregation strategies and interaction between visual and semantic features

We compare the method with and without semantic features in Table 1 of the main paper. Below, in Table 3 we show other aggregation options such as averaging and dot product. All results are using ED as semantic embedding.

Method	HMDB51
Average	33.1 ± 2.9
Dot Product	33.9 ± 3.2
Weighted Average	35.3 ± 3.6
Concatenation	43.2 ± 1.9

Table 3. Results on different aggregation options for the semantic and visual embeddings.

6 Seen and Unseen Class Performance for GZSL

In order to better analyze performance of the model on GZSL, we report the average seen and unseen accuracies along with their harmonic mean. The results using different embeddings and on the UCF101, HMDB51 and Olympics datasets are reported in Table 4. The reported results are on the same splits for fair comparison [2].

Model	E	Olympics			HMDB51			UCF-101		
		u	s	H	u	s	H	u	s	H
WGAN [4]	A	50.8	71.4	59.4	-	-	-	30.4	83.6	44.6
OD [3]	A	61.8	71.1	66.1	-	-	-	36.2	76.1	49.1
CLASTER	A	66.2	71.7	68.8	-	-	-	40.2	69.4	50.9
WGAN [4]	W	35.4	65.6	46.0	23.1	55.1	32.5	20.6	73.9	32.2
OD [3]	W	41.3	72.5	52.6	25.9	55.8	35.4	25.3	74.1	37.7
CLASTER	W	49.2	71.1	58.1	35.5	52.8	42.4	30.4	68.9	42.1
WGAN [4]	S	36.1	66.2	46.7	28.6	57.8	38.2	27.5	74.7	40.2
OD [3]	S	42.9	73.5	54.1	33.4	57.8	42.3	32.7	75.9	45.7
CLASTER	S	49.9	71.3	58.7	42.7	53.2	47.4	36.9	69.8	48.3
CLASTER	C	66.8	71.6	69.1	43.7	53.3	48.0	40.8	69.3	51.3

Table 4. Seen and unseen accuracies for CLASTER on different datasets using different embeddings. 'E' corresponds to the type of embedding used, wherein 'A', 'W', 'S' and 'C' refers to manual annotations, word2vec, sen2vec and combination of the embeddings respectively. 'u', 's' and 'H' corresponds to average unseen accuracy, average seen accuracy and the harmonic mean of the two. All the reported results are on the same splits.

References

1. Brattoli, B., Tighe, J., Zhdanov, F., Perona, P., Chalupka, K.: Rethinking zero-shot video classification: End-to-end training for realistic applications. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4613–4623 (2020)
2. Gowda, S.N., Sevilla-Lara, L., Kim, K., Keller, F., Rohrbach, M.: A new split for evaluating true zero-shot action recognition. arXiv preprint arXiv:2107.13029 (2021)
3. Mandal, D., Narayan, S., Dwivedi, S.K., Gupta, V., Ahmed, S., Khan, F.S., Shao, L.: Out-of-distribution detection for generalized zero-shot action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9985–9993 (2019)
4. Xian, Y., Lorenz, T., Schiele, B., Akata, Z.: Feature generating networks for zero-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5542–5551 (2018)